

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique



Université d'Oran
Faculté des Lettres, Langues et Arts
Département de Traduction

MEMOIRE
Présenté en vue de l'obtention du diplôme de
MAGISTER

Par
KADI ALLAH Fayçal

Thème de recherche

**LA TRADUCTION AUTOMATIQUE : état de l'art et les
problèmes inhérents (Arabe-Français-Anglais)**

Date de soutenance : 04/10/2011
Devant la commission d'examen :

Président:	M. Daoud	Professeur (Université Oran)
Rapporteur:	A. SEDDIKI	Professeur (Université Oran)
Examineur:	W. Tehami	Professeur (Université Oran)
Examineur	A. Zaoui	Maitre de Conférences (Université Oran)

ANNEE2010/ 2011

SOMMAIRE

INTRODUCTION.....	1
-------------------	---

CHAPITRE I : Historique

I-Introduction.....	5
- 1 ^{ière} étape.....	5
- 2 ^{ième} étape.....	10
- 3 ^{ième} étape.....	11
II- Les différents systèmes de T.A.....	13
III- Situation de la traduction automatique.....	15
IV- T.A et Intelligence Artificielle.....	16
V- T.A et traduction assistée.....	18
VI- Complexité de la T.A.....	20
VI- Conclusion.....	22
-Bibliographie.....	24

CHAPITRE II : Généralités sur la T.A

-Introduction.....	25
I- Les niveaux de traitement linguistiques.....	26
I-1- Approche à base de connaissances.....	26
I-1-a- Analyse morphologique et lexicographique.....	27
I-1-b- Analyse syntaxique.....	28
I-1-c- Analyse sémantique.....	29
I-1-d- Analyse pragmatique.....	29
I-1-e- Exemples de données.....	30
II- Interdépendance entre les niveaux.....	30
III- Exemples de difficultés.....	30

II-1- Les difficultés lexicales.....	30
IV- Les connaissances nécessaires pour la T.A.....	33
IV-1- Les dictionnaires.....	33
IV-2- Les connaissances syntaxiques.....	35
IV-3- Les connaissances sémantiques.....	36
IV-4- Les connaissances contextuelles.....	37
V- Les différents systèmes de T.A et types d'informations.....	38
V-1- Systèmes basés sur des règles.....	41
-La tendance lexicaliste.....	42
-Les lexiques et la génération des textes.....	49
V- 2- Systèmes basés sur un corpus de textes bilingues.....	50
V-3- Systèmes destinés aux utilisateurs particuliers	56
VI-4- Conclusion.....	57
Bibliographie.....	60

CHAPITRE III : Description de quelques systèmes de T.A existants

I-Le système ALT-J/E.....	62
I-1- Les dictionnaires.....	62
I-2- Le schéma général du système.....	63
I-3- Fondement théorique du système.....	63
I-4- Mécanisme de la traduction du système.....	64
II- Le système Ariane.....	65
II-1- Le processus de traduction.....	66
II-1-1- Analyse.....	67
II-1-2- Transfert.....	67
II-1-3- Génération.....	67
II-2- Evolution d'Ariane.....	68
II-3- Utilisation de l'aspect multilingue d'Ariane.....	69

II-4- Application - traducteur de l'anglais vers l'arabe.....	69
a- Développement de transfert vers l'arabe.....	69
b- Développement de la génération de l'arabe.....	70
III-le système Aleth Trad.....	71
III-1- l'analyse des textes.....	71
III-2- Traduction, révision et validation.....	72
III-3- Traduction multi-niveaux.....	73
III-4- Architecture d'Aleth Trad.....	75
IV- Le projet METAL.....	77
IV-1- Historique et évolution de METAL.....	77
IV-2- Le système METAL.....	77
IV-3- Les outils du développement lexical.....	80
IV-4- Evolutions.....	82
V- le système TOSHIBA.....	83
V-1- Introduction.....	83
V-2- Les caractéristiques du système.....	84
V-3- Configuration du système.....	84
V-4- La méthode de traduction.....	85
V-4-1- Analyse morphologique.....	85
V-4-2- Analyse syntaxique.....	86
V-4-3- Analyse sémantique.....	88
V-4-4- Transfert structural.....	89
V-4-5- La Génération.....	91
VI- Le système TORJOMAN.....	93
VI-1- Introduction.....	93
VI-2- Aspects fonctionnels.....	94
VI-3- Aspects techniques.....	94
VI-3-1- La compilation des dictionnaires.....	94

VI-3-2- L'analyse de l'Anglais.....	96
VI-3-3- Le transfert et la génération.....	97
VI-4- Caractéristiques du système TORJOMAN.....	97
VI-5- Architecture du moteur de traduction.....	98
VI-6- Remarques sur le système TORJOMAN.....	99
VII- Le système Systran.....	100
VII-1- Le système.....	100
-L'analyse.....	100
- La synthèse.....	101
- transfert.....	101
Conclusion.....	106
Bibliographie.....	108

CHAPITRE IV : Quelques difficultés liées à la T.A

I-Délimitation des unités de traduction	111
I-1- le mot est-il une unité valable ?.....	112
I-2- Classification contextuelle.....	121
I-3- Classification cohésionnelle.....	121
II-Conception des dictionnaires électroniques internes.....	125
II-1- L'étiquetage.....	126
II-2- La modélisation.....	126
II-3- Observations sur les besoins en dictionnaires.....	127
II-4- Exemple : les dictionnaires du système systran.....	132
Bibliographie.....	135

CHAPITRE V : La T.A et la langue arabe

I- Introduction.....	137
I-1- La théorie du mot graphique en arabe.....	138
I-1-1-Relations contextuelles et formants de mots.....	139

I-1-2-Le schéma de la grammaire des formants du mot	140
I-1-3- L'évolution des modèles de représentation des phénomènes au niveau du mot en Arabe.....	141
I-1-4- Structure de l'unité lexicale simple en Arabe.....	143
I-2- Les spécificateurs morphosyntaxiques du mot.....	148
I-2-1- Les 2 types de M-spécificateurs.....	148
I-2-2- Le schéma des spécificateurs morphosyntaxiques du mot	149
I-3- Conclusion.....	155
II-T.A et le verbe arabe.....	157
II-1- Introduction.....	157
II-2- Problématique de la structure verbale arabe.....	157
II-3- Conclusion.....	163
III- Exemples des difficultés de T.A des verbes anglais vers l'arabe	163
III-1- Introduction.....	163
III-2-Problématique de la T.A des temps des verbes anglais vers l'Arabe.....	165
III-3- Conclusion.....	172
Bibliographie.....	173
Conclusion général	176

Introduction

La traduction automatique (T.A) qui fait partie d'un domaine plus large, le traitement automatique des langues (TAL), a pour objectif de traiter des données linguistiques (textes) exprimées dans une langue dite "naturelle", avec pour particularité de transférer ces données dans une autre langue (naturelle). Ceci signifie que la T. A n'est autre que le transfert du 'fond' et de la 'forme' d'un texte, pris comme représentant de la langue de départ, d'un système linguistique vers un autre système linguistique avec tout ce que cela implique comme somme d'informations et de connaissances à 'projeter' vers ce système linguistique à partir d'un système linguistique donné de départ.

L'objectif ou la finalité de la T.A reste, donc, la conception de logiciels ou programmes, capables de traiter de façon automatique des données linguistiques, c'est-à-dire des données exprimées dans une langue (dite "naturelle"). Ces données linguistiques peuvent être des textes écrits, ou bien des dialogues écrits, ou encore des unités linguistiques de taille inférieure à ce que l'on appelle habituellement des textes (par exemple : des phrases, des énoncés, des groupes de mots ou simplement des mots isolés).

Le traitement, dit automatique, par opposition à un traitement manuel ou instrumental opéré par l'humain, utilise un ordinateur c'est-à-dire une machine conçue pour effectuer des calculs. Un traitement automatique est une suite d'actions ou calculs à faire effectuer par la machine dans un certain ordre chronologique, c'est-à-dire un programme. 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.

La T.A met en œuvre des outils et des techniques qui sont de trois ordres :

- linguistiques
- formels
- informatiques

L'élaboration de systèmes performants passe donc par des recherches fondamentales, en matière notamment de compréhension de texte et de génération de texte. Dans ces deux perspectives, le traitement de la langue porte non seulement, comme mentionné précédemment, sur les formes, mais aussi sur le contenu ; il doit mettre en œuvre 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. De telles recherches revêtent nécessairement un caractère pluridisciplinaire, et doivent associer étroitement linguistes et informaticiens.

Dans ce sens, si les aspects théoriques ou méthodologiques ont été largement privilégiés jusqu'ici dans les recherches portant sur la traduction, les ressources linguistiques ont été, en revanche, largement négligées. Cette mauvaise appréciation de la situation est responsable du retard accumulé dans le domaine, car les données doivent être suffisamment représentatives et non prototypales. On ne peut imaginer un système de traduction qui ne prendrait en compte que quelques centaines de mots et structures d'énoncés comme ce fut le cas dans beaucoup de programmes qui ont sous-estimé l'importance des ressources linguistiques. Dès lors, on a pris conscience d'investir massivement non seulement dans le développement des méthodologies mais également dans celui des ressources linguistiques.

Cela passe par des modèles et données linguistiques cohérents et explicites qui prennent en charge le langage dans sa totalité et non pas de manière

fragmentaire qui a souvent mené à des impasses, des contresens, ou même à un ‘dévoiement’ de sens, et donc, à des traductions erronées.

La construction de modèles fiables reste ardue, mais réalisable en tenant compte de la spécificité de chaque langue : le général et le particulier.

Les modèles linguistiques et informatiques existants sont complexes et difficilement applicable, si ce n’est inapplicable, à la traduction. Le développement de modèles adaptés exige le développement d’outils fondamentaux qui sont:

- la mise au point de méthodes et de formalismes prenant en compte des réalités plus complexes que l’identification de mots ou de structures syntaxiques ; il s’agit en particulier de prendre en compte la représentation du sens au niveau des énoncés et des textes ainsi que de leurs caractéristiques pragmatiques ; les méthodes développés doivent être applicables à différentes langues et créer ainsi un cadre général pour le processus de traduction automatique.
- L’utilisation de connaissances spécifiques pour cerner mieux l’interprétation des énoncés.
- La description exhaustive des structures syntagmatiques et syntaxiques.

Ceci n’empêche pas que la traduction totalement automatique de qualité demeure pour l’instant une perspective à long terme, pour laquelle des efforts importants doivent encore être consentis.

Dans ce travail, on tente de faire une compilation des données linguistiques, de réunir l’ensemble des cas de traduction qui demeurent ‘indigeste’ quant à un traitement automatique. Des cas qui seront puisés des langues françaises et anglaises vers la langue arabe. On procédera ensuite par donner des solutions susceptibles de surmonter ces aléas. Par exemple, la délimitation de l’unité linguistique ‘traduisible’ par un traitement automatique, que ce soit en Anglais ou en Français, reste toujours imperceptible alors que celle-ci

constitue la pierre angulaire de la T.A dans le sens où elle permet d'annihiler le mot en tant qu'unité linguistique et ainsi éviter le leurre sémantique qu'il constitue. Ou encore, laquelle des unités : linguistiques, sémantiques ou bien de traduction, qui est doit être prépondérante et donc privilégiée lors du traitement automatique ? Serait-ce toujours le cas quelle que soit la typologie ou type de texte ? Quel modèle, et donc représentation, approprié faut-il choisir pour une analyse adéquate du texte à tous les niveaux que ce soit morphologique, syntaxique, sémantique, pragmatique... ? Le type d'informations à extraire d'un texte est-il commun et reproductible ? Comment allier les connaissances linguistiques et les informations extralinguistiques ?...

C'est ce genre de questions auxquelles on tentera de donner des réponses en donnant un large éventail des travaux faits dans ce sens et les limites de chacun d'eux....

Avant cela, on énumérera les différents niveaux de traitement, les différentes approches et des exemples de modélisation.

On préconisera, au fur et à mesure, des solutions en tant que suggestions sachant, de notre part, que le sujet demeure complexe et un travail individuel reste en deçà de la consistance de la matière à traiter et le 'gigantisme' du sujet...

Mon approche ne sera pas ici de nature technique pure, ressortissant au domaine de l'informatique, ni même proprement théorique, ressortissant au domaine purement linguistique, elle est essentiellement pragmatique, dans la mesure où je m'intéresserai à un certain type de difficultés de traduction.

Chapitre I : Historique

I : Introduction

On peut 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'enthousiasme des débuts, puis 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.

- 1ère étape :

Les deux premiers précurseurs de la traduction automatique sont Andrew D. Booth et Warren Weaver. Mais c'est à W. Weaver (mathématicien américain), vers la fin des années 40, que revient en premier l'idée d'instruire la traduction dite automatique en proposant d'utiliser les techniques du déchiffrement cryptographique. En effet, influencé par les théories de Shannon, ce dernier considéra la T.A comme une sorte de cryptographie auquel un simple décodage suffira à transférer en langue cible. Un intérêt stratégique énorme était en jeu : dans le contexte d'après-guerre, la possibilité de traduire automatiquement des textes intéresse au plus haut point l'armée et les services d'espionnage. Les recherches commencent dans l'euphorie : les langues traitées se limitent au russe et à l'anglais. Booth, quant à lui, commença ses tentatives dès les années 47, en collaboration avec Richens, à qui l'on doit le principe de segmentation des mots en radicaux et désinences. Un mémorandum publié par Weaver en 1949 finira par

lancer la T.A. En 1952, le M.I.T organisait la première conférence dans le domaine.

L'ampleur que prirent les recherches était surtout, donc, motivée par le contexte de rivalité entre les Etats-Unis et l'URSS (guerre froide). C'est ainsi que les premiers systèmes pilotes aux Etats-Unis, tel que le G.A.T (Georgetown Automatic Translation), étaient consacrés à la traduction des textes scientifiques du russe vers l'anglais.

Les espoirs nés de ces premières recherches furent de courte durée puisqu'on s'est vite aperçu de leurs limites. On avait très largement sous-estimé les difficultés théoriques d'une pareille entreprise. Ainsi dès le début des années 60, on s'interrogea sur la faisabilité de traductions entièrement automatiques de bonne qualité.

En 1951, Yeoshua Bar-Hillel, dans son rapport, stipula l'impossibilité d'une traduction purement automatique et la nécessité d'une intervention humaine en amont (préparation du texte) et en aval (révision), la répartition du traitement en morphologie et syntaxe 'opérationnelle', la difficulté d'élaborer une grammaire universelle sur la base des travaux des logiciens et sur ceux de Harris ou de lui-même, la notion de 'grammaires de transferts'...il conclut d'ores et déjà qu'il était impossible d'imaginer une traduction entièrement automatique de bonne qualité sans prendre en considération le sens, un problème qui semble insoluble à l'époque.

Ce désaveu n'est pas surprenant et l'échec des recherches en T.A qui suivirent trouve de nombreuses explications, entre autres, l'informatique en est à ses premiers balbutiements (les mémoires sont limitées et les programmes ne peuvent pas être distingués des données linguistiques), les méthodes de programmation sont primitives ; ensuite la linguistique est peu développée : le structuralisme qui fonde son étude sur des énoncés

réalisés et ne s'intéresse pas aux problèmes du sens n'est pas directement utile pour les systèmes généraux de traduction automatique de l'époque ; enfin les chercheurs ne se rendaient pas compte des problèmes que pose la traduction.

Les systèmes de cette période, appelés d'ailleurs les systèmes de 'première génération', étaient basés sur des traductions mot à mot à partir d'un dictionnaire sans construire de représentation syntaxique de la phrase adoptant une approche dite directe. Ils ne procédaient pas à une analyse de la langue source séparée de la génération de la langue cible. En d'autres termes, le traitement linguistique était absent lors de l'élaboration de ces systèmes ce qui explique son échec.

En général, Un système de ce type est bilingue et fonctionne grâce à plusieurs dictionnaires et à un ensemble de programmes. Il n'est fondé ni sur des théories linguistiques ni sur des théories de compilation des langages formels.

Le processus de traduction consiste essentiellement en :

- 1-le remplacement des mots de la LS par ceux de la LC selon les dictionnaires,
- 2-quelques arrangements locaux par des programmes ad-hoc, L'aspect purement bilingue de ces systèmes exige :

- le recensement des phénomènes linguistiques les plus intéressants sur un couple de langue,
- la codification de ces phénomènes sous forme de dictionnaires et de programmes.

Les recherches menées à la même époque en URSS présentaient, par contre, une tendance nettement plus théorique puisqu'elles intégraient les questions d'analyse linguistique. Elles s'orientaient déjà vers

l'élaboration d'un 'langage intermédiaire' universel, indépendant de la langue source et de la langue cible, en termes de modélisations logico-mathématiques. On peut citer dans cet ordre là les travaux de l'équipe de V.Rosenzweig à Moscou, à laquelle ont participé des linguistes de renom comme I.Mel'chouk. Ces systemes préluèrent déjà à des systemes ultérieurs dits de 'deuxième génération'.

En 1965, la commission ALPAC (Automatic Language Processing Advisory Committee) présentait le rapport qui lui avait été demandé par les agences gouvernementales finançant depuis dix ans aux États-Unis les travaux de la vingtaine d'équipes de recherche en traduction automatique. Celui-ci fut défavorable au regard des résultats des traductions russe-anglais des systèmes américains de première génération par rapport à la qualité et le coût de ces traductions qui nécessitaient d'importantes révisions par l'humain. Il prônait la nécessité de développer les recherches en linguistique informatique condition, qu'il considérait, nécessaire pour progresser dans le domaine.

Les chercheurs se tournèrent donc vers deux directions différentes :

- des recherches théoriques en syntaxe : à la fin des années 60, les théories de Noam Chomsky sur les grammaires formelles et la grammaire générative transformationnelle ont un double impact. Tout d'abord, Noam Chomsky montre que la structure du langage est assez régulière pour être décrite de manière formelle dans une grammaire qui rend compte de la compétence linguistique des êtres humains . En second lieu, il développe une théorie qui débouche sur un traitement purement syntaxique du langage : elle présuppose en effet que la syntaxe est indispensable pour extraire le sens de la phrase et que l'analyse syntaxique peut se faire indépendamment de toute analyse du sens.

- Cette conception soulève un certain nombre de questions : est-il vraiment possible de faire une analyse syntaxique non ambiguë et générale d'une phrase, sans prendre en considération son sens, les connaissances du monde ou le contexte d'utilisation ? pouvons-nous, par exemple, donner une représentation non ambiguë de la phrase sans considérations sémantiques ou pragmatiques ? l'analyse syntaxique est-elle vraiment indispensable pour extraire le sens des phrases ? dans ce contexte, différentes études se tournent vers le problème de la représentation des connaissances en intelligence artificielle (IA). Ces recherches aboutissent à des résultats concrets dans le domaine du traitement des langues naturelles et à un certain nombre de programmes qui vont témoigner de la possibilité de traiter automatiquement le langage naturel, sans mettre nécessairement en œuvre une théorie linguistique, ce qui soulève différents problèmes.

Ainsi la T.A devint la locomotive de la linguistique formelle et de la linguistique computationnelle ou informatique. Au MIT, Yngve puis Chomsky travaillent sur la formalisation de la syntaxe.

Le travail des nombreuses équipes engendra une diversification des méthodes. Si la traduction directe continua à dominer de par son utilisation, deux nouvelles approches voient le jour dans la foulée : le système interlangue et le système par transfert.

Ses inconvénients :

- l'accumulation des phénomènes linguistiques devient si importantes que l'insertion de nouveaux phénomènes dans les programmes, de plus en plus complexe, devient pratiquement impossible. Le niveau de qualité atteint assez vite ses limites.
- les programmes linguistiques sont directement écrits dans les langages de programmation de bas niveau.

- **2^{ième} étape :**

Elle s'étalait sur la période allant de 1965 à 1975. C'est une période qu'on peut considérer comme peu prolifique mais néanmoins novatrice dans la vision qu'elle prônait sur la traduction automatique. Ainsi la plupart des systèmes commençaient à s'orienter vers ce qu'on a appelé les systèmes de 'deuxième génération'.

Les systèmes de deuxième génération avaient comme principales caractéristiques :

- une approche dite indirecte 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 météorologique). Le vocabulaire et les constructions sont ainsi plus facilement répertoriés, et les risques d'ambiguïtés moindres.
- Une diversification des langues traduites en dissociant le travail sur la langue source et sur la langue cible, ce qui permet d'étendre les systèmes à une pluralité de langues.

Parmi ces systèmes, on cite, aux Etats-Unis, METAL (à l'université de 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- (Grenoble).

Ainsi, les systèmes de la deuxième génération permettent la séparation des données linguistiques (les dictionnaires et les grammaires) et les programmes (compilateurs et moteurs des transducteurs).

Cette séparation permet d'utiliser le même outil logiciel pour toutes les langues et la division du processus en trois phases est bien adaptée à la traduction multilingues.

Ses inconvénients:

- Dans un système de deuxième génération la sémantique n'est exprimée que par des traits sémantiques qui sont analogues à des traits grammaticaux.
- Le problème des ambiguïtés est souvent difficile à résoudre, la cause est le manque de critères linguistiques ou sémantiques pertinents.
- **3^{ième} étape** :

Elle débute en 1975 à nos jours et constitue une reprise en matière de travaux dans le domaine de la traduction automatique. Ce regain d'intérêt est motivé à la fois par les besoins croissants des entreprises privées et publiques en matière de traduction, et par la sophistication 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 sont, pour la plupart, des systèmes de ‘deuxième génération’, plus perfectionnés que les précédents ; en particulier certains, comme ARIANE par exemple, sont capables d’un certain auto-contrôle. Ainsi grâce à des langages informatiques évolués permettant d’appeler récursivement et de façon contrôlée des sous-grammaire, ils peuvent détecter un cheminement erroné, revenir en arrière, entreprendre un autre chemin, 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 début des années 80 a été marqué par le lancement de plusieurs grands projets de traduction par ordinateur : en France le projet national de traduction assistée par ordinateur (PN-TAO, 1983-87), 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, et au Japon , le vaste projet de traduction et d’interprétation (traduction automatique, systèmes de visualisation multilingues, banques de données terminologiques...) qui annonce déjà une ‘troisième’ et une ‘quatrième’ génération de systèmes de traduction, qui devraient être implantés sur des ordinateurs de ‘cinquième génération’.

Le principe des systèmes de troisième génération est de viser une compréhension explicite du texte source:

- 1-Le texte source est analysé et transformé en une représentation conceptuelle indépendante de la langue, il s'agit d'une interprétation (au sens logique) dans un modèle formalisé.
- 2-Divers mécanismes d'inférence utilisant des connaissances extra-linguistiques et celle du contexte ont permis d'enrichir la représentation conceptuelle .
- 3-A partir de la structure obtenue, on produit le texte cible.

Les études et les réalisations actuelles ne sont que des approches partielles de certains aspects du problème, et cela pour des textes concernant des domaines très restreints.

II- les différents systèmes de traduction automatique :

Le travail de nombreuses équipes aboutit à une diversification des méthodes. Si la traduction directe – d’abord conçue comme un mot-à-mot à peine amélioré, puis intégrant une part variable d’analyse syntaxique- domine largement, s’ajoutent deux autres approches. D’où la typologie suivante, qui perdurera.

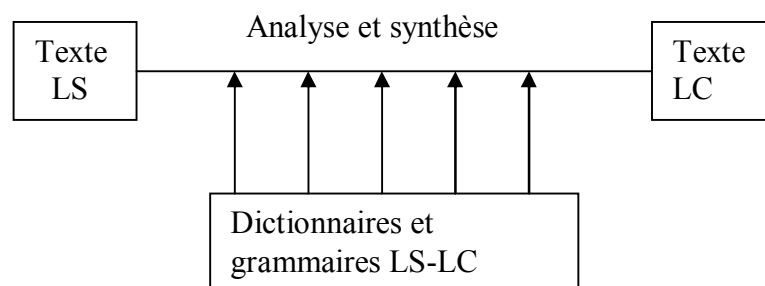


figure1 : Système de traduction direct

Un système direct n'opère que les analyses indispensables pour la traduction d'une langue dans une autre : l'analyse de la syntaxe et du vocabulaire de la langue source n'est pas poussée au-delà de ce qui est indispensable pour choisir les expressions appropriées et l'ordre des mots de la langue (figure1).

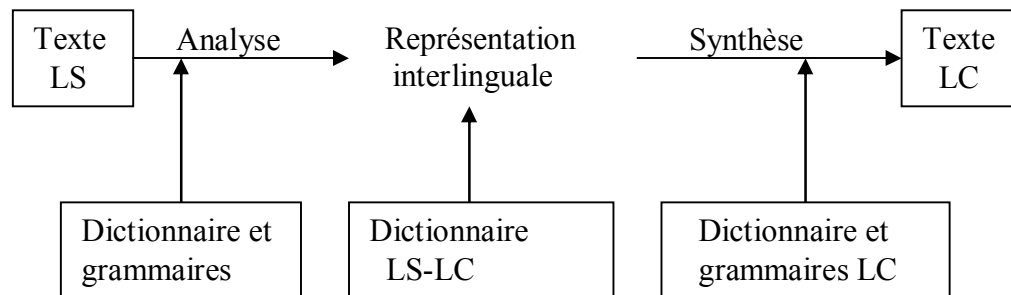


figure2 : Système interlingual

Un système interlingual procède en deux étapes nettement séparées : l'analyse complète du texte source en une représentation interlinguale et la synthèse du texte cible à partir de celle-ci (figure2).

Dans un contexte de multitraduction, cette option est théoriquement beaucoup plus économique : si l'on doit traduire dans n langages, il faut $n(n-1)$ programmes de traduction directe et seulement $2n$ programmes interlinguaux. L'interlingual utilisée peut être une langue logique artificielle, une langue semi-naturelle comme l'espéranto, un ensemble de primitives sémantiques universelles...

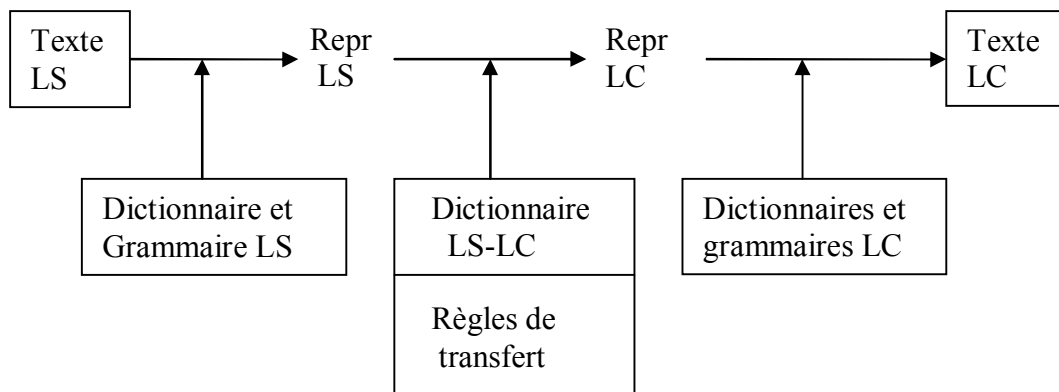


figure3 : Système par transfert

On appelle système par transfert un système qui comporte trois étapes, la première transforme le texte source en représentations de transfert de la langue source, la seconde transforme celles-ci en représentation de transfert de la langue cible, à partir desquelles est opérée la synthèse du texte en langue cible.

III- Situation de la traduction automatique :

A partir de 1975, on a vu apparaître un grand nombre de systèmes opérationnels et commerciaux tels que Systran, Logos, Météo et d'autres. Ils étaient généralement basés soit sur la méthode 'directe' ou sur la méthode de transfert syntaxique. Ils se reposent sur des dictionnaires bilingues assez riches. L'analyse syntaxique n'était pas profonde et il n'y avait presque aucune analyse sémantique, et l'exploitation des connaissances non-linguistiques était tout à fait absente.

La recherche était, jusqu'à la fin des années quatre-vingt, basée sur les règles linguistiques : les règles d'analyse syntaxique, les règles lexicales, les règles pour la formation de représentations abstraites, les règles de désambiguïsation, les règles de transformation des arbres syntaxiques, les

règles de transfert lexical, les règles de génération syntaxiques et morphologiques, etc.

A partir d'environ 1985, on a commencé à développer des systèmes basés sur les connaissances du domaine des textes à traduire, mais cette approche est restée une nouveauté presque jusqu'à la fin de la décennie.

Depuis 1989 ce cadre dominant a été rompu par l'entrée en scène de méthodes et de stratégies nouvelles qu'on appelle maintenant les méthodes "basées sur corpus". En premier lieu, un groupe à IBM a publié en 1989 les résultats de ses expériences avec un système de traduction purement statistique. L'efficacité de cette méthode a surpris beaucoup de chercheurs et a inspiré l'expérimentation avec les méthodes statistiques dans les années suivantes. En second lieu, certains groupes japonais ont commencé en même temps précisément à publier leurs résultats préliminaires avec des méthodes basées sur un corpus d'exemples de traductions. La caractéristique principale des deux approches est que les textes sont comparés et les traductions lexicales sont choisies sans utiliser aucunes règles syntaxiques ou sémantiques.

IV- Traduction automatique (T.A) et intelligence artificielle (I.A):

Le lien existant entre la T.A et l'I.A ne s'est avéré inéluctable qu'en phase des systèmes de 2^{ième} génération. Dans ce sens que l'I.A s'intéressait aussi à la compréhension automatique de textes. En effet avant, les deux domaines étaient étrangers l'un de 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. Elaboration de maquettes expérimentales de compréhension, où la traduction joue surtout un rôle de test).

L'intérêt des chercheurs dans ce domaine se portait principalement, du point de vue théorique, sur le mécanisme relationnel entre sémantique, pragmatique et connaissances d'univers (comme par exemple les travaux de T.Winograd ou R.Schank).

Dans cette perspective ont été construits des petits systèmes de compréhension limités à des domaines très restreints, et susceptibles de s'intégrer dans des systèmes de T.A.

Ainsi 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. Une telle représentation était conçue comme jouant le rôle de 'langage-intermédiaire' dans un système de traduction dénommé MAGIE.

Y.Wilks à l'université de Stanford a, lui aussi, 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.

On peut aussi évoquer 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', et le système MOPTRANS qui, de son côté, utilisait la notion de 'MOPs' ('memory organization packets').

Les systèmes de 'troisième génération', illustrés notamment par le système MU au Japon, tendent à rapprocher les recherches dans les deux domaines. 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 le biais de son seul sens linguistique.

V-Traduction automatique et traductions assistées :

Le terme de 'traduction automatique' regroupe en fait une grande variété de systèmes. L'expression 'traduction automatique' cherchait à qualifier au départ cet espoir d'arriver à une automatisation complète du processus de traduction.

A strictement parler, la T.A 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 pris du terrain sur celle de 'traduction automatique'.

En effet, selon l'importance du travail qui incombe respectivement à l'homme et à la machine, on peut distinguer deux types de traduction 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, et la traduction humaine assistée par l'ordinateur ('machine aided human translation'), qui fait jouer un moindre rôle à la machine ; dans un système de ce type, 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.).

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 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 parfois à la dénomination originelle de traduction 'automatique'.

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 ‘traductique automatique’.

VI- Complexité de la traduction automatique :

La traduction automatique est sans doute l’une des applications les plus complexes du domaine du traitement des langues naturelles (TLN) et ce, pour deux raisons : si l’on admet l’importance des connaissances linguistiques, elle nécessite au moins l’analyse du texte à traduire, indispensable à l’extraction du sens, et la production d’un texte porteur du même sens en langue cible. Or, aucune de ces deux étapes n’est triviale.

Le problème de l’extraction du sens du texte source suscite en effet de nombreuses questions théoriques, qui ne sont pas souvent discutées par les concepteurs de systèmes et sont loin d’être résolues : Qu’entends-on par le sens d’un texte ? Comment le représenter ? Correspond-il à la somme du sens des phrases ? Le sens des phrases peut-il à son tour être dérivé du sens des mots ? Quelles informations sont pertinentes pour la traduction ? la représentation doit-elle être commune à plusieurs langues (comme le soutient l’approche interlangue) ou spécifique à une langue précise (conformément à l’idée de l’approche basée sur le transfert) ? Est-il toujours nécessaire d’arriver à une représentation sémantique non ambiguë ?

Il est vrai que ces questions théoriques sont souvent négligées pour des considérations plus pratiques : comment extraire une représentation non ambiguë ? Quelles connaissances sont indispensables à cette extraction ? Comment stocker ces connaissances et les utiliser lors du processus de traduction ?

L'analyse des langues naturelles se heurte en effet au problème de l'ambiguïté des langues naturelles, ambiguïté à la fois lexicale et structurale. On parle d'ambiguïté lexicale quand un mot peut recevoir plusieurs catégories ou significations. Ce terme englobe aussi les ambiguïtés bilingues qui désignent deux mots qui, tout en étant la traduction l'un de l'autre, ne dénotent pas exactement le même concept ou le même ensemble de concepts. Le terme d'ambiguïté structurale fait référence à une phrase qui peut être dotée de plusieurs représentations. Dans certains cas, ces ambiguïtés peuvent être considérées comme telles par un être humain et sont dites réelles ; dans d'autres, elles sont inhérentes au traitement informatique (ou accidentelles).

Si la résolution de toutes les ambiguïtés accidentelles n'est pas toujours nécessaire pour traiter les langues proches, elle est malheureusement indispensable pour traduire correctement des langues de familles plus éloignées.

Il est donc souvent nécessaire de choisir la bonne interprétation, ce qui nécessite des connaissances linguistiques, contextuelles, générales ou statistiques et même des capacités de raisonnement. Si les recherches ont beaucoup progressé dans ce domaine, la représentation et l'intégration de ces sources de connaissances restent encore souvent expérimentales et on se pose souvent la question de savoir jusqu'à quel point ce type d'informations est suffisant et adéquat pour résoudre les problèmes d'analyse.

Au niveau de la production du texte en LC, les problèmes sont tout aussi importants et sans doute encore moins abordés dans la littérature. On sait qu'il est possible d'exprimer le même sens de différentes manières en fonction du contexte linguistique ou pragmatique et un système devrait être capable de choisir, parmi toutes les variantes

stylistiques possibles, celle qui convient le mieux dans ce contexte spécifique.

VII- Conclusion :

L'histoire mouvementée de la T.A/ T.A.O soulève toute une série de questions. On s'aperçoit que le terme même de traduction recouvre toute une palette d'activités, répondant à des besoins diversifiés. Comme pour d'autres applications de l'informatique, la machine ne doit pas être conçue comme le rival de l'homme, utopie qui conduit à une impasse pratique, l'ordinateur n'est qu'un outil adapté à certains usages. Dans cette optique, la traduction de grande qualité complètement automatisée apparaît purement académique : en effet, l'automaticité ne doit pas exclure la révision, opération normale en traduction humaine ; quand à la qualité, elle dépend de ce que l'on veut faire du document. On doit placer au centre le point de vue des utilisateurs : seules deux choses leur importent : les systèmes peuvent-ils produire des documents de qualité suffisante pour l'usage visé ?

Boitet exprime un point de vue analogue : 'les systèmes de TA ne traduisent pas comme les traducteurs humains...la seule chose importante est que leur résultat soit une sorte de traduction acceptable par des humains pour certaines tâches, qui vont de l'acquisition de l'information à la révision professionnelle'.

Comme dans d'autres domaines des industries de la langue, la traduction met en lumière la complexité des langues naturelles et la nécessité de collecter une somme de connaissances linguistiques et pragmatiques considérables pour les traiter. La participation accrue des linguistes, associés à d'autres secteurs des sciences cognitives, est une des caractéristiques de la période. Mais si la linguistique

computationnelle prend la traduction comme un terrain d'expérimentation, elle doit du même coup se plier aux exigences industrielles et aux besoins des usagers.

BIBLIOGRAPHIE

ABBOU André, *observatoire francophone des industries de la langue (1989) Traduction Assistée par Ordinateur, Perspectives technologiques, industrielles et économiques envisageables à l'horizon 1990 - Actes du séminaire international (Paris,Mars 1988)*, Ed. DAICADIF, 1989. 234 p.

Bachut. D, *le projet EUROLANG : une nouvelle perspective pour les outils d'aide à la traduction*, 1994.

Bédard. C. , *la prétraduction automatique : outil de productivité et d'évolution professionnelle*, *Meta XXXVII*, 1992, p.738-760 .

TMI-90. Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, 11-13 June 1990, University of Texas, Austin, TX, USA

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

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

MT Summit III, July 1-4 1991, Washington D.C., USA.

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

Hutchins, W.J. (1993) Latest developments in machine translation technology. In: **MT Summit IV** (1993: 11-34)

Vasconcellos, M. (1993) The present state of machine translation usage technology, or: How do I use thee? Let me count the ways! In: **MT Summit IV** (1993: 35-46)

CHAPITRE II :GENERALITES SUR LA T.A

Introduction:

La tâche d'un système de traduction automatique peut être définie simplement comme suit:

L'ordinateur doit être capable, à partir d'un texte d'une **Langue Source (LS)** de produire son équivalent dans une **Langue Cible (LC)**.

La qualité du texte produit en génération LC est irréprochable et similaire au travail d'un traducteur professionnel.

Les efforts pour la conception et la réalisation des systèmes de TA peuvent être résumés par des solutions aux problèmes posés par le texte en LS ou en LC (ie. les diverses facettes de compréhension).

Un certain nombre de questions importantes peuvent être envisagées sur ce point de vue :

- 1- Quelle est la signification du texte ?
- 2- A-t-il certaines composantes structurelles ?
- 3- Quel est le représentant de la compréhension d'un texte ?
- 4- Quel est l'ensemble des extractions de significations d'un texte?
- 5- Est-il absolument nécessaire d'extraire la signification (tous les termes signifiants ou de compréhension) dans un but de traduire ?

La **question 1-** soulève un problème de base en linguistique et en philosophie du langage. Notre tâche est dévouée à la théorie et la méthodologie dans le champ de la linguistique et de l'**Intelligence Artificielle (IA)**.

La **question 2-** concerne les composantes structurelles qui nous intéressent sont d'ordre : logico-morphologique, lexicale, syntaxique, lexical-sémantique et contextuelle (sémantique inférentielle et pragmatique).

La **question 3-** se rapporte aux problèmes de la représentation des connaissances, aussi bien dans le style et le contexte fait en IA.

La **question 4-** concerne les reflets des problèmes computationnels, qui sont l'entreprise de la traduction automatique.

La **question 5-** implique les réponses négatives qui reflètent les souhaits des chercheurs en TA et les limites pratiques de l'état de l'art en TA ou TAO.

Ces questions soulèvent des problèmes difficiles à résoudre et pour le moment aucune solution définitive n'a été trouvée dans le développement en linguistique computationnelle et en IA .

I-LES NIVEAUX DE TRAITEMENT LINGUISTIQUE

Le traitement linguistique repose sur deux principes directeurs qui sont : l'approche à base de connaissances et la séparation de connaissances. L'usage des deux principes est apparu au début des années 90 (premier prototype de système de traduction à base de connaissances, Nirenburg, 1991) pour combler le traitement du sens auquel était privilégié le traitement des aspects formels de la structure syntaxique prôné par la grammaire générative et transformationnelle.

I-1- APPROCHE A BASE DE CONNAISSANCES :

La mise en œuvre d'une véritable analyse linguistique nécessite, donc, la capacité, pour le système de pouvoir :

- reconnaître (niveau morpholexical)

- structurer (niveau syntaxique)
- comprendre (niveau sémantique)
- contextualiser (niveau pragmatique)

I-1-a-ANALYSE MORPHOLOGIQUE ET LEXICOGRAPHIQUE :

Problèmes à résoudre :

- qu'est-ce qu'un mot ?
- comment segmenter un texte ?
- que se passe-t-il pour les langues sans séparateurs ?
- quels types de séparateurs faut-il prendre en compte ?
- comment traiter les formes inconnues : 'fichiers.bat', 'minitelisation'...

- la morphologie :

Les problèmes précités relèvent principalement de la morphologie (étude de la structure des mots) et de la lexicographie (recensement et classification des formes d'une langue). Ces domaines s'intéressent plus particulièrement à la dimension paradigmatique de la langue (par opposition à sa dimension syntagmatique).

Pour illustrer les problèmes abordés ici, nous reprenons l'exemple de (Delafosse 1999), soit la chaîne de caractères :

Jean a mangé des pommes

La segmentation se fera de la manière suivante :

U1= Jean, U2= a mangé, U3= des, U4= pommes, U5=. (point)

Maintenant on peut associer toutes sortes d'informations aux U_i ($i=1,2,3,\dots$), comme par exemple :

U1 = Jean :

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

U2 = a mangé

- forme lemmatisée : manger
- informations morpho-syntaxiques : verbe, passé composé, indicatif, 3^{ème} personne, singulier, construction : transitif, ...

Idem pour U3 et U4,...

I-1-b-ANALYSE SYNTAXIQUE :

La syntaxe vise à l'étude des contraintes entre les catégories morphosyntaxiques devant être prises en compte pour la description des séquences de mots 'acceptables' dans une langue donnée. Les contraintes peuvent être de nature sélectionnelle (règles d'accord) ou positionnelle. La description des contraintes caractéristiques d'une langue se fait par le biais d'une grammaire.

La syntaxe permet de résoudre (ou réduire) des ambiguïtés aux niveaux inférieurs :

- au niveau phonétique :

il mange

ile m'en jeu

.....

- au niveau lexical :

il dst parti

il est dit ...

Elle peut aussi faciliter la formulation des connaissances sémantiques ou pragmatiques. Par exemple, elle impose des contraintes sélectionnelles pour le verbe 'manger' : sujet animé, objet comestible.

I-1-c- ANALYSE SEMANTIQUE :

Au sens littéral, la sémantique vise à l'étude du sens hors contexte.

Le niveau sémantique est encore beaucoup plus complexe à décrire et à formaliser que les niveaux de traitement précédents, par conséquent les réalisations qui sont opérationnelles sont peu nombreuses, et elles concernent des applications très limitées où l'analyse sémantique se réduit à 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 traitement sémantique prend comme unité d'analyse la phrase, et conduit à représenter sa partie significative. Ces 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, que le sens résulte de la double donnée du sens des mots et du sens des relations entre mots.

I-1-d-ANALYSE PRAGMATIQUE :

La pragmatique vise à l'étude du sens en contexte.

L'analyse sémantique de la phrase isolée, traitée hors contexte, ne conduit à représenter que la partie de la signification des mots dans cette 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. C'est la raison pour laquelle une analyse pragmatique est nécessaire, et qui consiste à trouver la signification 'réelle' des phrases liées aux conditions situationnelles et contextuelles d'utilisation des mots.

I-1-e-EXEMPLES DE DONNEES :

- au niveau morpholexical : règles morphologiques (grammaire du mot) et construction de lexiques informatiques
- au niveau syntaxique : la construction de grammaires (formelles) de la langue
- aux niveaux sémantique et pragmatique : construction de modèles formels des connaissances sur le monde (propositions logiques, réseaux sémantiques, graphes conceptuels,...)

II-INTERDEPENDANCES ENTRE LES NIVEAUX :

- nécessité de connaissances syntaxiques :
- nécessité de connaissances sémantiques
- nécessité de connaissances pragmatiques
- interdépendance syntaxe-sémantique

Il mange du poisson avec une fourchette (nécessité de connaissances sémantiques).

La petite brise la glace (nécessité de connaissances pragmatiques).

III-EXEMPLES DE DIFFICULTES :

- difficultés au niveau lexical :
- problème de la correction lexicale (mots erronés) :

les formes erronées ne sont pas toujours évidentes à repérer : ‘la lasse du chien’

III-1-Les difficultés lexicales :

Elles découlent de la conception particulière du ‘mot’ par l’ordinateur puisqu’il est perçu comme une suite de caractères encadrée de séparateurs (espace blanc ou signe de ponctuation). Sachant qu’il existe dans des langues comme le français ou l’anglais par exemple des expressions composées telles que pomme

de terre, rouge-gorge, pied-à-terre, hard-ass, hard copy ..., il devient difficile à l'ordinateur de distinguer ces mots composés comme 'mot' unique. L'unique solution réside au recours à la traduction dite compositionnelle c'est à dire répertorier tous les mots composés et leurs équivalents dans la langue cible, chose qui n'est pas évidente quand on sait qu'il en existe déjà environ 2 millions (français et anglais) et la prolifération de nouveaux autres exclusivement techniques tels que floppy disk, virus-ridden...et la liste n'est pas close... !!! Il y a aussi dans le même registre les expressions verbales composées comme porter aux nus, crier fort, to feel hard done ...et toutes les formes idiomatiques anglaises. La première tâche à effectuer donc est de répertorier ou stocker toutes les expressions composées et savoir reconnaître les expressions verbales dans les phrases parce qu'elles ne sont pas forcément contiguës comme le montre l'exemple suivant :

- Il fut porté aux nues.
- Tous les présents ont porté cette vedette aux nues.

On trouve aussi des difficultés homographiques :

- de catégories de discours différents (nom, verbe, préposition...)
 - le garçon (article défini) ; je le vois (pronom)
 - Il marche rapidement (verbe) ; la marche est le meilleur sport (nom)
- de même catégorie mais de sens différent (polysémie):
 - Il mange un avocat (fruit) ; Il a rencontré un avocat (homme de loi)
 - la marche est programmée pour demain (manifestation) ; la marche est bénéfique à la santé (sport) ; tu montes la marche (bâtiment).
 - his eye hurts him (organe) ; to have a good eye (flair); the eye of the needle (couture).

Ce sont là quelques ambiguïtés lexicales introduites par l'ordinateur dont l'identification sémantique n'est nullement perceptible à moins d'user d'algorithmes ou méthodes de raisonnement basées sur des règles linguistiques (grammaire ...) ou sur des connaissances du monde.

En guise d'illustration d'algorithmes utilisant les règles linguistiques pour levée l'ambiguïté lorsqu'il s'agit de catégories de discours différents:

Il marche rapidement

(il) (marche) (rapidement)

il=pronom, marche=nom

l'ordinateur se pose la question

marche=nom/verbe ?

levée de l'ambiguïté

il=pronom, pronom.nom interdite, pronom. Verbe autorisé

Donc marche=verbe

En ce qui concerne les ambiguïtés de même catégorie mais de sens différent, le meilleur moyen étant d'utiliser la catégorisation c'est à dire la classification sémantique de toutes les catégories grammaticales. Ensuite on définit toutes les relations conditionnelles entre ces catégories afin de 'détecter' ou 'délimiter' le champ sémantique du mot. Ainsi si on considère les deux exemples suivants :

- l'élève saisit la leçon.
- L'élève saisit le stylo.

L'ordinateur ne peut trancher entre les deux homographes 'saisit' dans le sens de 'comprendre' dans le premier et 'saisit' dans le sens de 'tenir' dans le second pour pouvoir opter pour la traduction appropriée de ces deux mots. La procédure de délimitation du champ sémantique serait par exemple comme suit :

Classification sémantique des verbes :

Saisir1 = classe de 'action mentale'

Saisir2 = classe de 'action physique'

Classification ou catégorisation des compléments du verbe :

Leçon = classe 'abstrait'

Stylo = classe 'concret'

Relation conditionnelle :

Dire que pour ‘saisir1’ c’est à dire appartenant à la classe de ‘action mentale’ le complément d’objet doit appartenir à la classe ‘abstrait’, sinon c’est à dire si le complément appartient à la classe ‘concret’ l’ordinateur optera pour ‘saisir2’ c’est à dire classe ‘action physique’.

Donc pour la levée de l’ambiguïté lexicale concernant la polysémie, il faut procéder par catégorie et ensuite faire une interaction fonctionnelle et conditionnelle entre les catégories. Le problème qui se pose dans ce cas c’est l’intersection entre les catégories ou classes et la prolifération des combinaisons si on tient compte du fait qu’il peut exister des ‘interférences’ entre elles en plus de la multitude inimaginable de classes qu’on peut concevoir. Les relations conditionnelles génèrent des graphes et arborescences qui restent difficile à gérer.

On peut citer d’autres ambiguïtés comme les ambiguïtés croisées et les ambiguïtés pragmatiques.

IV. Les connaissances nécessaires pour la TA:

IV.1. Les dictionnaires:

Les traducteurs humains utilisent les dictionnaires comme source d'information pour les langues LS et LC.

Le type de dictionnaire souvent utilisé par l'homme est le dictionnaire bilingue , qui connecte les unités (mots) de la LS et la LC. Un point important à remarquer est que ces dictionnaires bilingues sont conçus pour l'usage humain.

La situation est peu différente quand le dictionnaire est utilisé par le programme d'un ordinateur.

On peut illustrer ces types de dictionnaire et les modules d'exploitation, comme suit: on suppose que le système reçoit en entrée un texte français: "le livre est sur la table" (1) .

La traduction en anglais de (1) est : "The book is on the table." (2). Le dictionnaire nécessaire pour la performance de cette traduction est le suivant :

<i>Allemand</i>	<i>Anglais</i>
sur	on
livre	book
le.....	the
la.....	the
est	is
table	table

Le programme de traduction a fourni une substitution des mots français en anglais un par un.

L'analyseur à ce niveau est **lexical** ou **orthographique**, ie: fondé sur l'aspect des mots ou aspect **morphologique**.

Peut-on utiliser le même système pour passer de l'Anglais vers le Français ?

Non , parce que dans ce cas plusieurs relations vont du mot "the" à "le" et "la".

Les connaissances inscrites dans ce dictionnaire sont insuffisantes pour résoudre cette ambiguïté et il faut ajouter d'autres connaissances pour faire le choix adéquat.

IV.2. Les connaissances syntaxiques :

L'un des problèmes de l'analyse automatique que résout la syntaxe est le processus de désambiguïsation relevant des différentes catégories d'un mot.

La syntaxe décrit la structure des mots et des connexions associées dans une phrase. L'analyseur à ce niveau est **grammatical**, fondé sur les règles d'agencement des groupes de mots et des mots entre eux.

Comme par exemple : "The coach lost a set." , sans connaissance de la structure de la phrase anglaise , il est impossible de savoir si :

- 1- le mot "coach" est un nom ou un verbe ,
- 2- le mot "lost" est un verbe ou un adjectif ,
- 3- le mot "set" est un nom ou un verbe ou un adjectif .

Ainsi les connaissances syntaxiques de la langue anglaise , nous permettent, dans ce cas, d'éliminer douze (12) cas d'ambiguïtés et de faire le choix correct pour la compréhension.

Pour ce faire , la reconnaissance de la structure syntaxique de la phrase par des règles de réécriture permet d'identifier les différents mots, unités ou groupes de la phrase (syntagmes nominaux , syntagmes verbaux...).

Ce type de connaissance est enregistré dans la grammaire anglaise ou dans toute autre grammaire de la LS d'un système de TA.

Les règles de réécritures pour notre exemple :

S ^NP + VP
NP D + N
VP V + NP

avec: S : (Sentence) phrase,
 NP : (Noun Phrase) Syntagme
 nominal, VP : (Verbal Phrase)
 Syntagme verbal, D : (Determiner)
 déterminant, N : (Noun) nom et
 V: (Verb) verbe.
 Le résultat de l'analyse est :

K S[NP[D(the),N(coach)],VP[V(lost),NP[D(a),N(set)]]].

IV.3. Les connaissances sémantiques:

La sémantique s'occupe du signifié des mots et des connexions associées dans une phrase. L'analyseur est fondé sur **des liens de signification entre les mots**, comme les anaphores, et traite les problèmes de polysémies (ie. les mots ayant plusieurs significations). L'une des premières tentatives dans le domaine de l'analyse sémantique est d'utiliser les représentations des connaissances d'un langage naturel pour un système de traduction. Ceci a été réalisé à l'université de YALE aux USA par les travaux de Shank en 1975: les Dépendances Conceptuelles (DC) dans un langage de représentation des connaissances .

Les DC sont utilisées pour représenter le sens de la phrase; les dictionnaires bilingues utilisés dans ces systèmes prennent en considération des choix de marques ou de mailles, pour représenter les connaissances sémantiques.

Ces mailles sont utilisées lors de la phase de génération pour que le système connecte la signification au mot adéquat dans la LC.

On illustre la représentation en DC d'une phrase anglaise :

phrase: "Mary hit John." DC-
 phrase : (event EV001
 (action PROPEL)

(agent Mary)
(object John)
(instrument *?* unknown)
(force * above-average *)
(intentionality * positive *)).

IV.4. Les connaissances contextuelles:

La pragmatique s'occupe de la cohérence générale des éléments sémantiques entre eux. L'analyseur est fondé sur le bon sens.

Exemple: « Alors que Marie parlait, Jean buvait ses paroles. »

Morphologie: tous les mots sont corrects du point de vue orthographique.

Syntaxe: les règles de grammaires sont correctement appliquées.

Sémantique: l'anaphore « ses » se rapporte à Marie, ie: Alors que Marie parlait, Jean buvait les paroles de Marie.

Pragmatique: le verbe « boire » implique un complément d'objet direct « liquide » or « les paroles » ne sont pas un « liquide » donc il y a une erreur pragmatique ! .

Par contre, si l'on sait que le contexte est « poétique », alors il s'agit d'une métaphore et la phrase est correcte. Les connaissances contextuelles permettent d'éliminer des ambiguïtés provenant de causes très diverses:

a) L'ambiguïté peut venir de la grammaire :

L'ambiguïté peut venir de la grammaire qui n'a pas pu choisir entre deux analyses. On illustre ce cas, par l'exemple suivant:

"Je viens d'écrire un article sur mon bureau."

Le syntagme prépositionnel "sur mon bureau", peut se rattacher à :

- 1- " article " ou
- 2- " écrire ".

b) L'ambiguïté peut venir du sens des mots :

Dans la phrase : "J'achète un terrain à mon voisin.", le verbe "acheter" exprime que le sujet reçoit la possession du complément direct qui était auparavant la possession du groupe prépositionnel.

c) La pragmatique ne peut pas tout résoudre:

Dans la phrase : " Pierre a acheté un bijou à ma voisine.", nous avons ici non seulement deux événements peu probables mais aussi des connaissances générales.

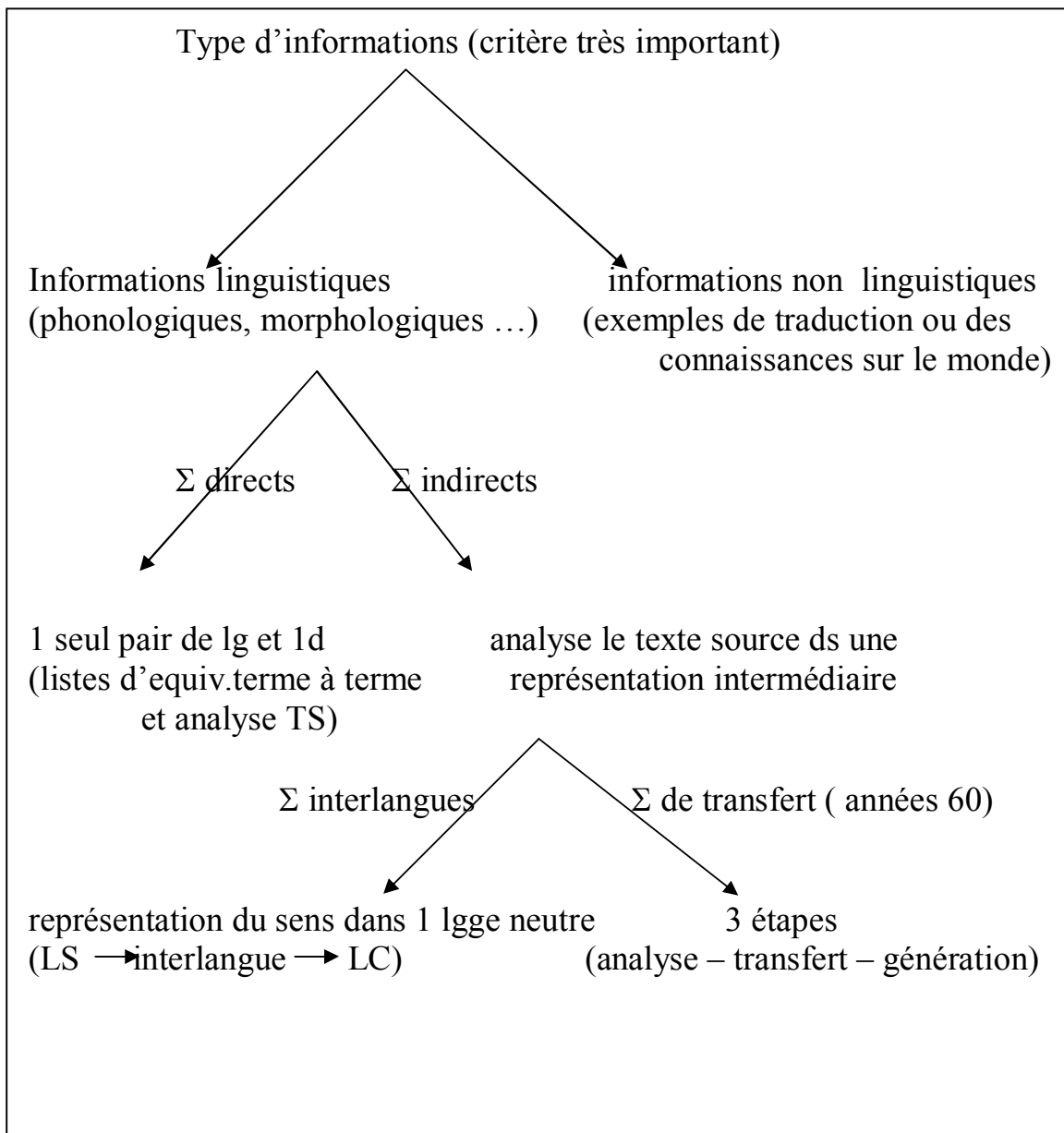
V- les différents systèmes de T.A et types d'informations :

La T.A ou traductique, qui est l'usage de l'informatique dans le processus de traduction dans le but d'automatiser le transfert linguistique d'un texte, pris comme 'échantillon' ou représentant, de la langue source vers la langue cible, ne comprend bien évidemment pas la T.H.A.O (traduction humaine assistée par ordinateur) et peut être scindée en deux grands domaines qui sont d'une part la T.A.A.H (traduction automatique assistée par l'homme) ou T.A.O (traduction assistée par ordinateur) ou encore T.A.A (traduction automatique assistée), et d'autre part la T.E.A.H.Q (traduction entièrement automatique de haute qualité) qui reste un objectif bien éloigné.

Les premiers systèmes reposaient sur la traduction mot à mot en procédant en 3 étapes : la lemmatisation, le transfert lexical de la langue source vers la langue cible et enfin un réajustement morpho-syntaxique. Si ce procédé simpliste prenait en charge certains aspects de la langue tel que l'aspect paradigmatique, il ne rendait toutefois pas compte de sa complexité. Et c'est tout naturellement

qu'on s'aperçut vite de ses limites. Ainsi on ne pouvait traduire mot à mot les mots composés, les mots polysémiques ... sans altérer le sens et aboutir à des déconvenues criardes. L'outil de traduction automatique ou ordinateur avait par conséquent besoin de comprendre. Qu'est ce que ça veut dire donc comprendre pour un ordinateur ? c'est tout simplement, et tout à la fois conceptualiser, modéliser, représenter et générer : c'est à dire faire abstraction de la forme au détriment du contenu sémantique. La forme ou graphie n'étant utile que pour le traitement différentiel ou analogique entre les unités sémantiques pour l'extraction du sens. 'Bien comprendre' est donc lié à la capacité de l'ordinateur de calculer une représentation sémantique du texte source.

La modélisation nécessite le recours à deux types d'information (voir schéma ci-dessous) selon qu'on soit rationaliste ou empiriste. Les rationalistes prônent le recours aux informations dites linguistiques (phonologiques, morphologiques). Les empiristes quant à eux préfèrent les informations dites non linguistiques comme les exemples de traduction ou des connaissances sur le monde. Techniquement, il existe deux types de systèmes de T.A : les systèmes directs qui traitent une seule paire de langues utilisant des listes d'équivalence terme à terme avec au préalable une analyse du texte source, et les systèmes indirects qui procèdent par une analyse du texte source dans une représentation intermédiaire. Ces derniers peuvent être soit des systèmes à interlangue (représentation du sens dans un langage neutre) ou des systèmes de transfert procédant en trois étapes : analyse, transfert et génération.



Il serait alors légitime de se poser la question sur le degré d'efficacité de tels traducteurs électroniques. Pour répondre à une telle question on peut dire que l'efficacité est subordonnée ou tributaire à leurs aptitudes à surmonter les difficultés inhérentes à la T.A. Les difficultés ont alors deux origines. Ceux qui découlent des informations non linguistiques comme les difficultés probabilistiques ou les difficultés d'enregistrement de tous les exemples possibles. Et les difficultés engendrées par les informations linguistiques c'est à

dire des difficultés rencontrées en analyse et en génération partant du fait que la langue est une infinité d'ambiguïtés. Ainsi d'une part une forme linguistique dispose de plusieurs sens possibles (difficultés en analyse) et d'autre part un sens donné peut se présenter sous différentes formes linguistiques (difficultés en génération). Ce sont des difficultés lexicales. En plus l'interaction de ces formes linguistiques entre elles engendre ce qu'on appelle les ambiguïtés syntaxiques des lors que le recouvrement de leur champs sémantiques n'est pas toujours évident. On se contentera de donner quelques notions de difficultés d'ordre lexical accompagnées d'exemples pris indistinctement du Français ou de l'Anglais.

V-1- Systèmes basés sur des règles :

Il y a cinq ou six ans on a vu la fin de deux grands projets de l'approche "transfert": le projet Ariane à l'Université de Grenoble et le projet Eurotra des Communautés Européennes. Ces systèmes illustraient les traits typiques des systèmes dits de la "deuxième génération", c'est-à-dire: trois étapes d'analyse, de transfert et de synthèse; des processus d'analyse et de génération qui parcourent plusieurs niveaux distincts de morphologie, de syntaxe et de sémantique; des représentations d'interface assez abstraites en forme d'arbres étiquetés; l'utilisation des règles pour transformer des arbres d'une étape à l'autre; traitement par lots avec post-édition et sans aucune intervention humaine au cours de la traduction; et une absence presque totale d'information pragmatique et textuelle.

Cependant, les systèmes de transfert ont continué. Il y a, par exemple, le système commercial Métal, le projet LMT d'une équipe à IBM, et le projet multilingue Eurolang qui est en train d'être développé par la compagnie SITE en France avec la collaboration de la compagnie allemande Siemens-Nixdorf, et qui se base sur les expériences du projet Eurotra.

L'approche "interlingue" a continué, elle aussi, avec encore plus d'énergie. Les traits distinctifs de cette approche sont bien connus: une langue pivot neutre pour représenter le sens des textes (l'interlingue,) et des banques de connaissances dans le domaine des textes à traduire. A l'Université de Carnegie Mellon plusieurs modèles ont été développés, et en 1992 on annonça l'inauguration d'un projet collaboratif avec la compagnie Caterpillar, dont le but sera un système pour la traduction sans post-édition des manuels techniques dans le domaine des engins de terrassement.

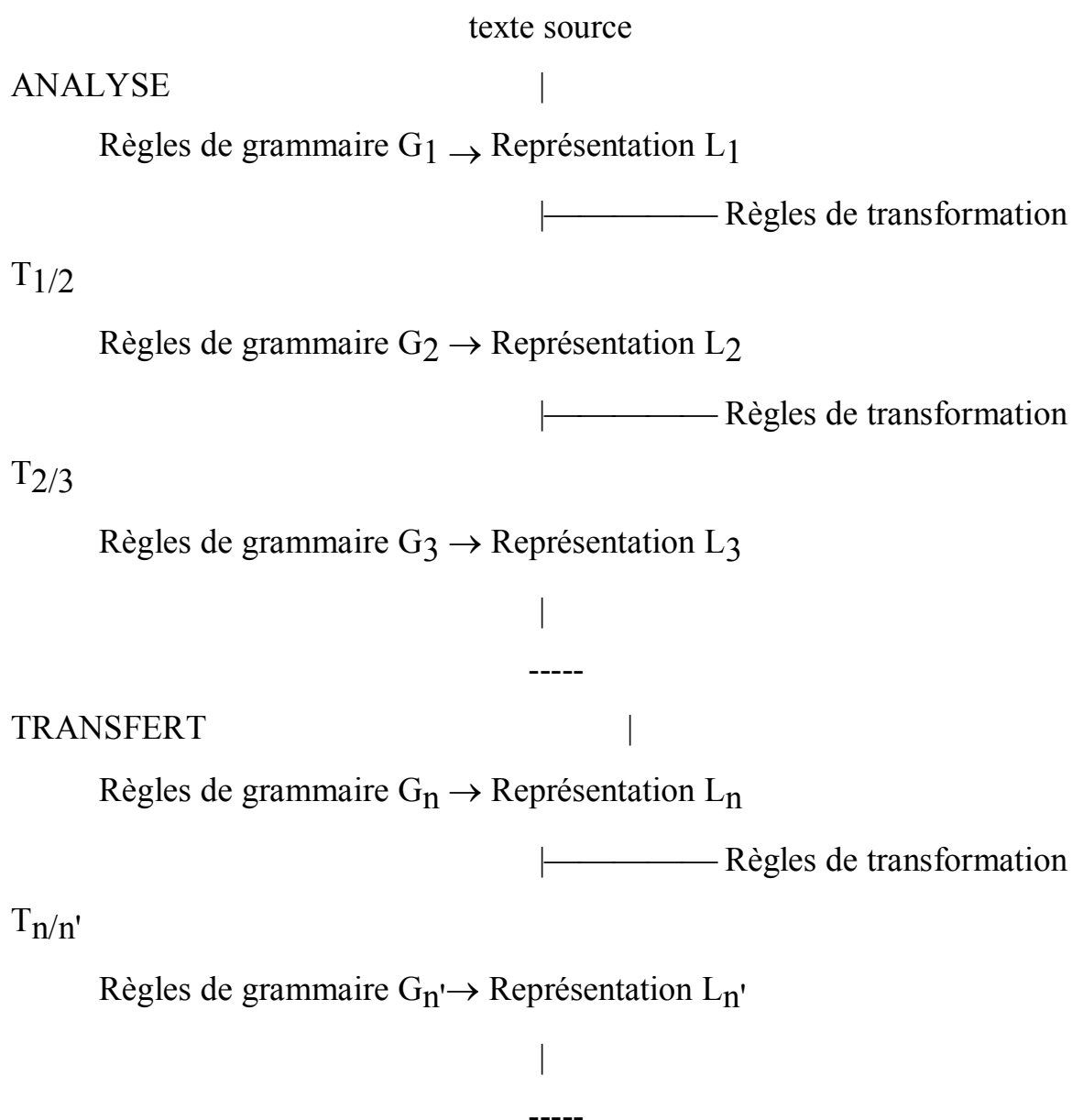
Il existe d'autres systèmes "interlingue", par exemple le système ULTRA à l'Université de New Mexico et le système UNITRAN basé sur la théorie linguistique des principes et des paramètres. On doit y ajouter le projet Pangloss, un système interlingue limité au sous-langage des fusionnements et des rachats. Celui-ci est un projet collaboratif des universités de Southern California, New Mexico State et Carnegie Mellon, et qui utilise les expériences de ces équipes dans leurs propres recherches.

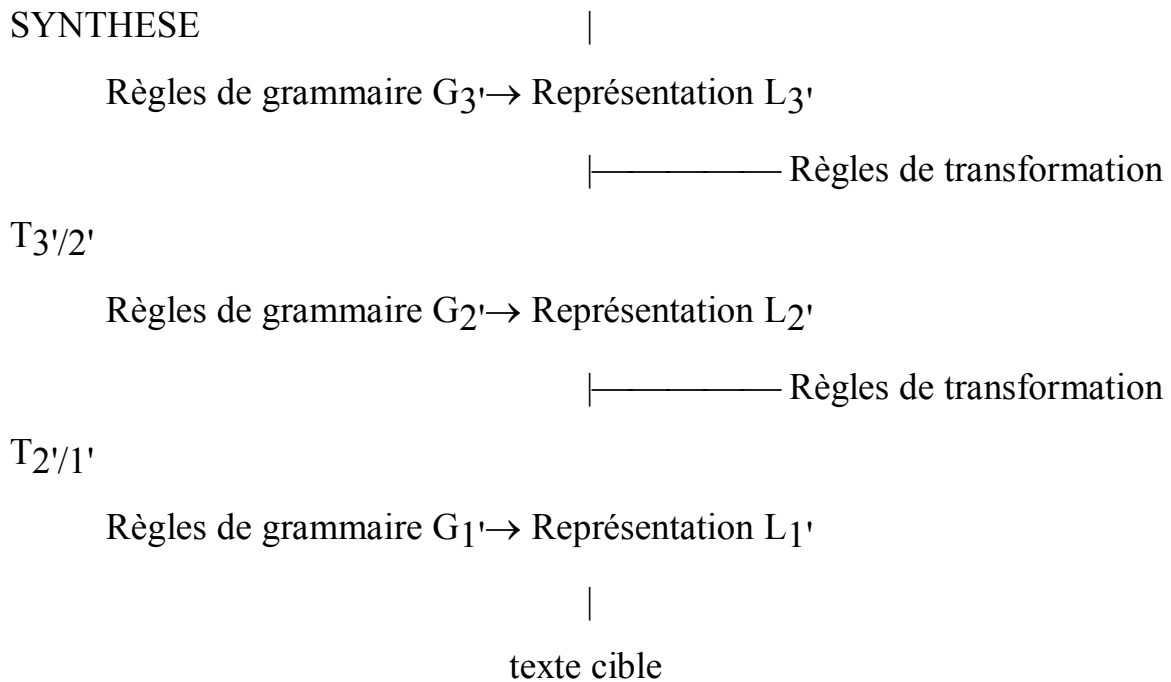
- la tendance "lexicaliste" :

Un trait caractéristique des systèmes basés sur les règles a été la transformation ou 'mappage' des représentations en forme d'arbres étiquetés. Par exemple *[fig.1]*, dans le système Eurotra on a proposé une série de transformations des arbres: un arbre morphologique est transformé en un arbre syntaxique, un arbre syntaxique en un arbre sémantique, un arbre d'interface du texte source en un arbre équivalent du texte cible, etc. Essentiellement, un arbre doit satisfaire des conditions précises, posséder une structure particulière et contenir des unités lexicales particulières ou des traits syntactiques ou sémantiques particuliers. En plus, les arbres eux-mêmes sont testés par les règles de formation - une

grammaire met à l'épreuve la structure et les relations qui y sont représentées. Un arbre est rejeté s'il ne conforme pas aux règles grammaticales du niveau en question: morphologie, syntaxe, sémantique, etc. Les grammaires et les règles de transformation déterminent les conditions ou contraintes qui limitent les possibilités de transfert d'un niveau à un autre et, en somme, d'un texte de la langue source à un texte de la langue cible.

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





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

Fig.2: Formalisme basé sur contraintes (LFG)

2 (a):

John likes Mary ↔ Marie plait à Jean

like, V:

(↑PRED) = like <SUBJ, OBJ>

(τ↑PRED FN) = plaire <SUBJ, OBJ>

(τ↑AOBJ OBJ) = (SUBJ)

(τ↑SUBJ) = (OBJ)

john, N:

(↑PRED) = john

(τ↑PRED FN) = jean

mary, N:

(↑PRED) = mary

(τ↑PRED FN) = marie

F-structure de la langue cible:

⌈ PRED	plaire	⌋	
SUBJ	[PRED marie]		
⌊ AOBJ	[OBJ	[PRED jean]	] ⌋

Student is likely to work ↔ Il est probable que l'étudiant travaillera

likely, A:

(↑PRED) = likely <XCOMP> SUBJ

(↑SUBJ) = (XCOMP SUBJ)

(τ↑PRED FN) = probable

(τ↑COMP) = (XCOMP)

probable, A:

(↑PRED) = probable <COMP>SUBJ

(↑SUBJ FORM) = il

(↑COMP COMPL) = que

F-structure de la langue cible:

	[	PRED		probable	]
		SUBJ		[FORM il]	
			[	PRED	travailler]
		COMP		COMPL que	
	[		[	SUBJ	[...]]]

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

Au lieu de l'orientation syntaxique qui a caractérisé beaucoup des systèmes de transfert dans le passé il existe actuellement une tendance vers les solutions lexicales. Un exemple extrême de l'approche "lexicaliste" est la méthode qu'on appelle en anglais "shake and bake" (en français "agiter et cuire") [fig.3].

Fig.3 Approche lexicaliste ('Shake-and-bake')

unité lexicale monolingue (anglaise):

		[	ORTHO		like	]
				[	like (E1),	]
[12]		SEM	E1:	{	role (E1, experier, X1),	}
				[	role (E1, stimulus, Y1)	]
		ARG0	E1			
		ARG1	X1			
	[	ARG2	Y1			]

unité lexicale monolingue (française):

	⌈	ORTHO		<i>aime-</i>		⌋
				⌈ aime (E2)		⌋
[13]		SEM	E2:	{ role (E2, stimulus, X2)	}	
				⌊ role (E2, experier, Y2)	⌋	
		ARG0	E2			
		ARG1	X2			
	⌊	ARG2	Y2			⌋

entrée lexicale bilingue pour *like-aime*:

⌈		⌈		⌈ ARG0 E ⌋⌋⌋
FRANCAIS[13]		SEM		ARG1 X
	⌊		⌊	ARG2 Y ⌋⌋
	⌈		⌈	ARG0 E ⌋⌋
ANGLAIS [12]		SEM		ARG1 Y
⌊	⌊		⌊	ARG2 X ⌋⌋⌋

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

Les origines des grammaires d'unification et des grammaires basées sur les contraintes remontent à presque dix ans d'ici. De nos jours, l'unification est

devenu un concept central pour un grand nombre de théories linguistiques, et les grammaires et les formalismes basés sur les contraintes attirent beaucoup de chercheurs en traduction automatique: LFG (Grammaire fonctionnelle lexicale), Definite Clause Grammar, GPSG, Grammaire catégoriale, etc. (Parmi les grammaires des contraintes on peut compter la grammaire d'UNITRAN basée sur les principes et les paramètres.) Le majeur avantage de ces grammaires est la simplification des règles d'analyse, de transformation et de génération. Au lieu d'une série de représentations complexes d'information à plusieurs niveaux, on voit des représentations mono-stratales ou même le transfert au moyen de unités lexicales simples. En même temps, les composantes de ces grammaires sont en principe réversible. Il n'est pas nécessaire de construire pour la même langue des grammaires différentes d'analyse et de génération; le même formalisme et les mêmes grammaires peuvent être appliqués dans les deux sens.

Plusieurs équipes ont construit des systèmes généraux pour le traitement de la langue naturelle, basés sur les grammaires d'unification et des contraintes, qui s'appliquent aussi à la traduction. Le système CLE (Core Language Engine), par exemple, a été appliqué à la traduction automatique du suédois à l'anglais et de l'anglais au suédois; tandis que le système PLNLP (Programming Language for Natural Language Processing) a fourni la base de systèmes de traduction pour le portugais, le koréen et le japonais. Mieux connu dans le domaine de la traduction automatique est l'Environnement Linguistique d'Unification (ELU) développé par une équipe suisse à Genève. Sur cette base on a fondé un système bi-directionnel pour la traduction des bulletins d'avalanches entre l'allemand et le français.

- les lexiques et la génération des textes :

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

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

Quant à la génération, il y a dix ans encore, on croyait que les composants qui posaient le plus de problèmes étaient ceux de l'analyse syntaxique et sémantique,

de la désambiguïsation du sens, de l'identification des antécédents des pronoms - en somme, de la compréhension du texte à traduire. A cette époque on négligeait pour la plupart les problèmes de la génération de textes idiomatiques dans la langue cible. Aujourd'hui, il est évident à tout le monde que, pour améliorer la qualité des traductions automatiques il faut non seulement faire attention à l'analyse mais aussi à la sélection des mots propres et à l'idiomatisme des textes produits. La génération de textes n'est plus un aspect négligé de la recherche en traduction automatique.

V-2- Systèmes basés sur un corpus de textes bilingues :

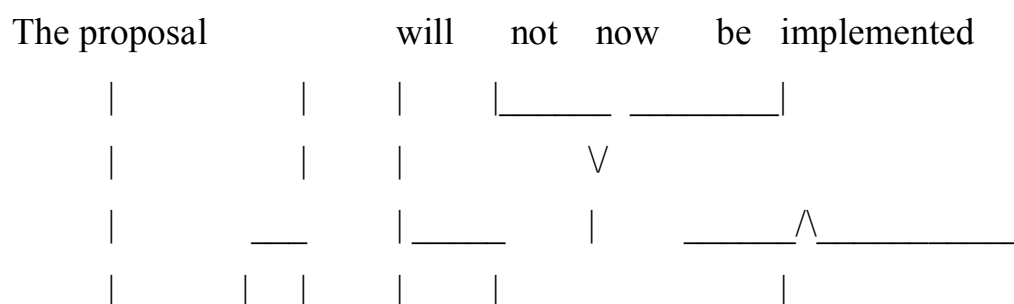
Ces deux tendances ont été accentuées fortement par les développements de nouvelles méthodes qui sont basées sur les corpus de textes bilingues, en particulier les méthodes statistiques et les méthodes basées sur des exemples de traductions.

Le développement le plus étonnant a été le projet Candide par un groupe de chercheurs à IBM. Sa méthode est fondée sur l'analyse stochastique de l'énorme corpus de textes en anglais et en français fourni par les rapports des débats parlementaires au Canada. La première tâche est d'aligner les phrases, les groupes de mots et les mots individuels des textes parallèles. Ce stage est achevé sans utiliser aucune information linguistique. On voit, par exemple [fig.4], que cette méthode a aligné *proposal* et *proposition*, *now* et *maintenant*, et le mot anglais *implemented* et la phrase française *mises en application*. Mais, contre l'intuition linguistique elle a aligné *will* et *seront*, tandis que pour le mot *be* aucun mot n'est aligné en français. A partir d'un grand nombre de tels alignements anglais-français on a calculé les fréquences et les probabilités des correspondances. Le mot anglais *not* correspond le plus souvent à deux mots français (fertilité 2 à une probabilité de 0.758), et ces deux mots sont en général *ne* et *pas* (probabilités de 0.469 et 0.460); les autres correspondances sont moins

probables: *non* (0.024), *pas du tout* (0.003), etc. La méthode a été évaluée par la traduction de l'anglais en français.

Fig. 4: TA stochastique (Candide, IBM)

Alignement:



Les propositions ne seront pas mises en application maintenant

Anglais: *not*

Français	Probabilité	Fertilité	Probabilité
<i>pas</i>	.469	2	.758
<i>ne</i>	.460	0	.133
<i>non</i>	.024	1	.106
<i>pas du tout</i>	.003		
<i>fau</i>	.003		
<i>plus</i>	.002		

etc.

Pour beaucoup de chercheurs, ce qui a été surprenant c'est que les résultats étaient tellement acceptables: presque la moitié des phrases traduites correspondait exactement aux traductions dans le corpus, ou bien le même sens était exprimé avec des mots un peu différents, ou bien d'autres traductions

légitimes étaient offertes. Pour améliorer ces résultats, les chercheurs d'IBM ont proposé d'introduire des méthodes statistiques plus complexes et aussi quelques éléments minimaux d'information linguistique -ils sont des empiristes et ils sont prêts à expérimenter avec d'autres moyens. Ils proposent d'utiliser l'information morphologique, p.ex. de traiter toutes les variantes d'un verbe comme un seul mot, et ils proposent aussi d'utiliser certaines transformations syntaxiques, p.ex. la transformation de structures discontinues en structures qui ressemblent plus exactement celles de la langue cible [fig.5]

Fig. 5: Transformations syntaxiques (proposées)

Has the store any eggs? → The store has any eggs QINV

John does not like turnips → John likes do_not_M1 turnips

Je ne sais pas → Je sais ne_pas

Je vous le donnerai → Je donnerai le_DPRO vous_IPRO

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

Fig. 6: Banque d'exemples des traductions: *field*

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

Quelques équipes utilisent les méthodes sémantiques, par exemple un réseau sémantique ou une hiérarchie de termes d'un domaine. D'autres équipes ont utilisé les informations statistiques à propos des fréquences lexicales dans la langue cible. Le plus souvent, la méthode des exemples est utilisée comme complément aux méthodes plus traditionnelles basées sur les règles linguistiques. Parce que les textes sont extraits des traductions humaines on a l'assurance que les résultats seront aussi idiomatiques que possible. C'est une des plus grandes difficultés de la traduction automatique du français à l'anglais de choisir la préposition propre pour le petit mot *de* [fig. 7]. Une banque de données qui offre un grand nombre d'exemples peut très bien aider la sélection des équivalents. Elle peut aider encore plus où il y des difficultés plus grandes, p.ex. pour la traduction en français de la phrase anglais *have an effect on* [fig. 8].

Fig. 7: *de*

français	anglais
----------	---------

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

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

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

Une banque de textes bilingue en parallèle peut être également utilisée dans d'autres buts. En particulier, plusieurs équipes conduisent les expériences avec des stations de travail destinées aux traducteurs humains. L'un des groupes les plus actifs dans les méthodes statistiques pour aligner les textes bilingues (c.a.d. le groupe à AT&T Bell) prévoit l'application d'un corpus aligné comme une base

de connaissances pour les traducteurs. Un autre groupe très actifs est l'équipe bien connue au Centre canadien de recherche sur l'informatisation du travail. Son but est le développement d'outils valables pour les traducteurs, y compris la facilité de chercher une banque de textes pour les exemples d'utilisation de n'importe quel mot anglais ou français dans un contexte particulier.

La méthode des exemples de textes a renforcé aussi la tendance dont j'ai fait déjà mention vers la recherche de génération de textes. En dehors de la traduction automatique, il y a des autres groupes qui s'intéressent aussi de plus en plus à la représentation des données informatiques dans une langue naturelle idiomatique. La génération de textes dans plusieurs langues a été le but de deux projets canadiens: dans le projet RAREAS c'est la production de textes anglais et français des prévisions maritimes; dans le projet LFS c'est la production de sommaires bilingues sur la marché du travail.

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

V-3- Systèmes destinés aux utilisateurs particuliers :

Il y a toujours eu des systèmes de traduction automatique qui sont largement limités à des domaines restreints. Même les systèmes destinés à l'usage général sont souvent effectivement limités à quelques domaines particuliers. La limitation diminue les difficultés de construire un lexique suffisant, et elle diminue les problèmes d'ambiguïté et de choix des équivalents dans la langue cible. Dans ces cas, il est possible de restreindre le vocabulaire et les structures grammaticales des textes à traduire. Bien que les frais de l'édition préalable peuvent être assez élevés, la post-édition sera réduite de façon considérable. On peut aussi limiter le système lui-même à un sous-langage spécifique. Un exemple bien connu est le système Météo qui traduit, depuis quinze ans, les rapports météorologiques. Parmi les systèmes de sous-langages et de langues restreintes des années plus récentes on compte le système CRITTER pour les rapports du marché des bestiaux, les projets déjà mentionnés (ELU, Pangloss, et Caterpillar), et les projets très ambitieux destinés au développement de systèmes pour traduire la langue parlée. Le projet ATR au Japon dure déjà depuis sept ans et continuera jusqu'à la fin du siècle; c'est un système pour les renseignements sur les conférences internationales et pour les enregistrements téléphoniques dans les hôtels. Le projet VERBMOBIL en Europe vise à développer une aide portative dans les négociations commerciales face à face conduites en anglais par les Allemands ou les Japonais qui ne connaissent pas très bien la langue anglaise.

Dans le passé il n'y a que peu de systèmes construits par les utilisateurs mêmes. Un exemple connu est la PAHO (Organisation Pan-Amérique de la Santé), où deux systèmes ont été développés pour la traduction de l'anglais à l'espagnol et de l'espagnol à l'anglais. Au cours des dernières années il y a eu plusieurs systèmes construits par les utilisateurs mêmes. Ce sont typiquement des

systèmes avec des vocabulaires restreints, pour un domaine particulier et basés sur un sous-langage spécifique. C'est un signe encourageant que les méthodes informatiques de la traduction automatique et du traitement de la langue naturelle se sont maintenant répandus de plus en plus au delà des cercles limités de chercheurs. De tels systèmes ne sont pas innovatifs au point de vue théorique et méthodologique, mais ils sont souvent d'une construction très avancée en ce qui concerne leur technologie informatique. Je crois que c'est une tendance que nous verrons se répandre rapidement au cours des années à venir.

V-4- Conclusion :

De l'avis de beaucoup des spécialistes les systèmes d'avenir combineront des méthodes de règles assez traditionnelles et des méthodes statistiques et basées sur les exemples. Ils seront hybrides. Mais de quelle façon? Dans une optique possible, on peut envisager que les méthodes linguistiques des systèmes indirects fourniront la base pour l'application des méthodes basées sur les banques de connaissances, sur les données statistiques et sur les exemples de textes traduits. La base linguistique fournira les analyses syntaxiques et sémantiques assez simples, l'essentiel du transfert et l'information syntactique pour la génération. A la base on verra probablement les formalismes d'unification et les grammaires basées sur contraintes. Les autres méthodes fourniront une désambiguïsation plus souple et les informations nécessaires pour le transfert lexical et pour la production des textes idiomatiques.

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

- les règles seront moins ambitieux que celles des systèmes indirects de transfert et d'interlingue;
- l'analyse syntaxique sera limitée à la reconnaissance des structures superficielles, des composantes de phrase et des relations de dépendance;
- il n'y aura presque aucune analyse profonde des relations logiques;

- l'analyse sémantique sera limitée à l'identification des rôles d'éléments dans les phrases: l'agent, l'instrument, etc.

- les informations lexicales seront extraites principalement des sources régulières comme des dictionnaires généraux destinés au grand public; par conséquence elles n'indiqueront que des catégories syntactiques et peut-être des traits sémantiques peu raffinés.

- ces traits sémantiques assez rudimentaires ne seront utilisés que pour une désambiguïsation initiale.

- les règles de transfert lexical et structural s'appliqueront peut-être à des représentations peu profondes (quoique moins simples que dans le système d'IBM).

Quant aux méthodes encore assez nouvelles:

- des exemples de traduction, stockés dans une banque de textes bilingues alignés, seront utilisés pour aider la désambiguïsation plus délicate dans l'analyse de la langue source et pour choisir les équivalents dans la langue cible

- des données statistiques sur les collocations lexicales et les fréquences du vocabulaire monolingues aideront l'analyse syntaxique et sémantique des phrases, la désambiguïsation monolingue, et le choix des phrases idiomatiques dans la langue cible

- des données sur les probabilités des équivalences bilingues seront utilisées pendant le transfert lexical

- des banques de connaissances des domaines en question aideront la désambiguïsation monolingue et interlingue

On peut anticiper aussi d'autres développements importants:

- l'emploi des méthodes de feedback pour améliorer les grammaires (ou le fond de règles) et les lexiques monolingues et bilingues

- les recherches plus approfondies sur les problèmes de discours et de style
- l'intégration de la traduction automatique dans les systèmes pour la production, la transmission et le management de documents dans les bureau (à l'exemple des postes de travail pour traducteurs)

BIBLIOGRAPHIE

ABBOU, André, MEYER, Thierry, LEFAUCHER, Isabelle (1987) : *Les industries de la langue. Les applications industrielles du traitement de la langue par les machines*, Paris, Éd. Daicadif, 400 p.

AUGER, Pierre (1988) : «L'industrialisation de la langue française et son maintien comme grande langue véhiculaire de la science et de la technique», 56^e Congrès de l'ACFAS, Moncton, 9 — 13 mai 1988, s.p.

BARANES, William. «Les industries de la langue», *Qui vive international*, n° 4, pp. 74-75.

BOULANGER, A., C. DUBUISSON, L. EMIRKANIAN (1985) : « Problèmes de catégorisation dans l'analyse automatique du discours », communication présentée au 53^e congrès de l'ACFAS tenu du 20 au 24 mai 1985 à l'Université du Québec à Chicoutimi (Chicoutimi).

CASSEN, Bernard (1986) : «Un nouveau front pour le français et la langue de l'Europe, LES INDUSTRIES DE LA LANGUE. ENJEUX POUR L'EUROPE», Actes du colloque de Tours, Tours, 28 fév. — 1^{er} mars 1986, Encrages, n° 16, nov. 1986, pp. 12-14.

CASSEN, Bernard, DEGREMONT, Jean-François (1986) : «Bilan de la Mission «industries de la langue» au 31 juillet 1986, LES INDUSTRIES DE LA LANGUE. ENJEUX POUR L'EUROPE. Actes du colloque de Tours, Tours, 28 fév. — 1^{er} mars 1986, Encrages, n° 16, nov. 1986, pp. 148-151.

CALSOLARI, Nicoletta (1984) : « Machine-Readable Dictionaries, Lexical

Date Bases and the Lexical System », *in Proceedings of Coling 84*, p. 460.

CERCONE, N. et G. McCALLA (1984): « Artificial Intelligence : Underlying Assumptions and Basic Objectives », *Journal of the American Society for Information Science*, vol. 35, n° 5, pp. 280-290.

DEGREMONT, Jean-François (1986): «Au croisement de l'informatique et de la linguistique, LES INDUSTRIES DE LA LANGUE. ENJEUX POUR L'EUROPE. Actes du colloque de Tours, Tours, 28 fév. — la-mars 1986, *Encrages*, n°16, nov. 1986, pp. 22-46.

FOX, M.S., D.J. BABEL, A.C. PARKER (1980) : « The Automated Dictionary », *Computer*, vol. 13, n° 7, pp. 35-48.

SLOCUM, Jonathan (1985): « A Survey of Machine Translation : Its History, Current Status, and Future Prospects », *Computational Linguistics*, vol. 11, n° 1, p. 17.

WEBER, Howard R. (1984): « Machine-Readable Components in a Variety of Information-System Applications », *in Proceedings of Coling 84*, p. 463.

CHAPITRE III: DESCRIPTION DE QUELQUES SYSTEMES DE T.A EXISTANTS

On dresse dans ce chapitre une liste non exhaustive de quelques systèmes de traduction. On présentera leurs caractéristiques générales. On essayera surtout d'axer sur ceux ayant trait à la langue arabe.

I-Le système ALT-J/E:

Il a été développé par le NTT, Laboratoire des Sciences de la Communication au Japon. Le projet a débuté en 1985.

I.1. Les dictionnaires:

Le système ALT-J/E est doté d'un grand dictionnaire sémantique:

	mots communs	noms propres	termes techniques	autres
Nombre d'entrées	100.000	200.000	70.000	30.000

Sachant qu'il y a 3000 catégories sémantiques qui sont ventilées sur 80 domaines.

- Un dictionnaire de transfert japonais/anglais de 16.000 structures verbales.
- Un dictionnaire de règles syntaxiques et sémantiques adapté au domaine d'application requis.

Celui-ci permet de prendre en compte le contexte au niveau d'un paragraphe, de traduire des expressions idiomatiques, de réécrire automatiquement des phrases japonaises en phrases standards, de traduire spécifiquement les enclitiques japonais, de traiter les oppositions, de générer des nombres et de comptabiliser les noms anglais, de générer des pronoms possessifs en anglais et enfin d'ordonner correctement les mots pour l'écriture des adresses.

I.2 Le schéma général du système: (Fig.III.1)

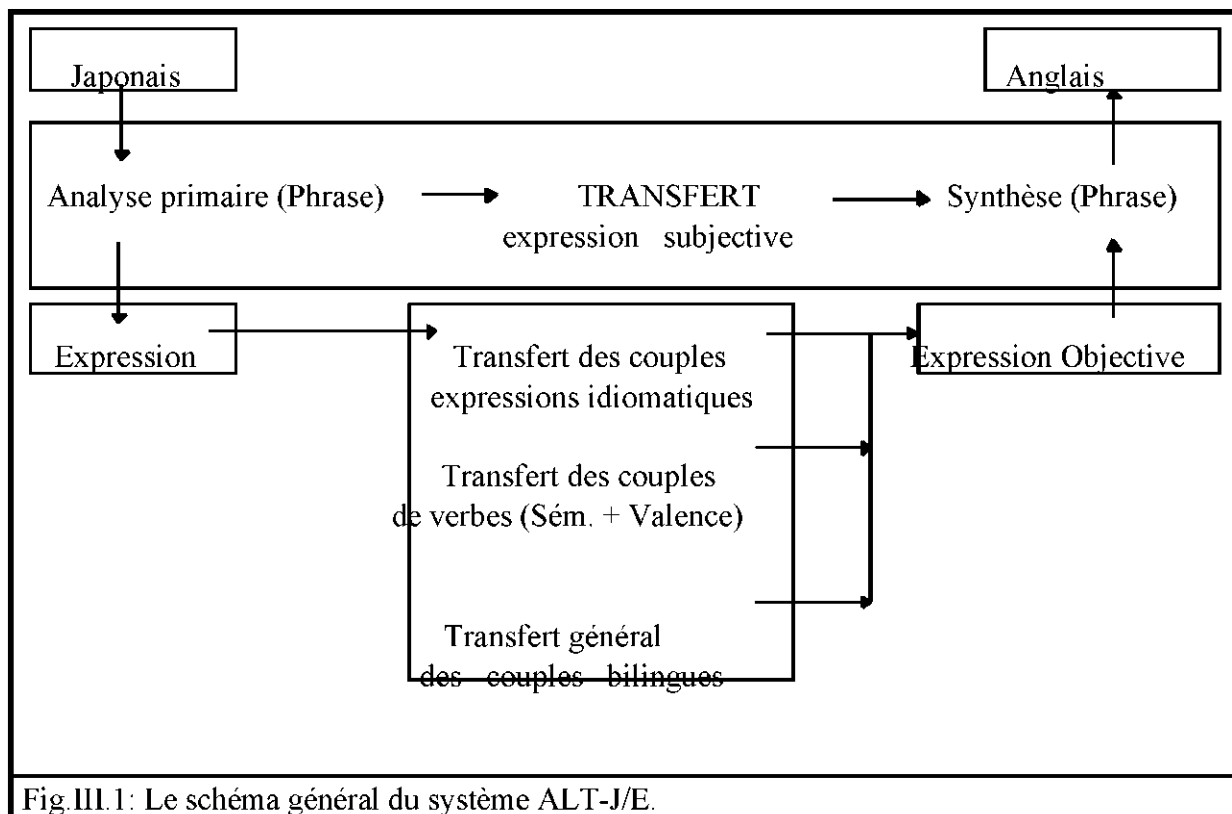


Fig.III.1: Le schéma général du système ALT-J/E.

I.3. Fondement théorique du système:

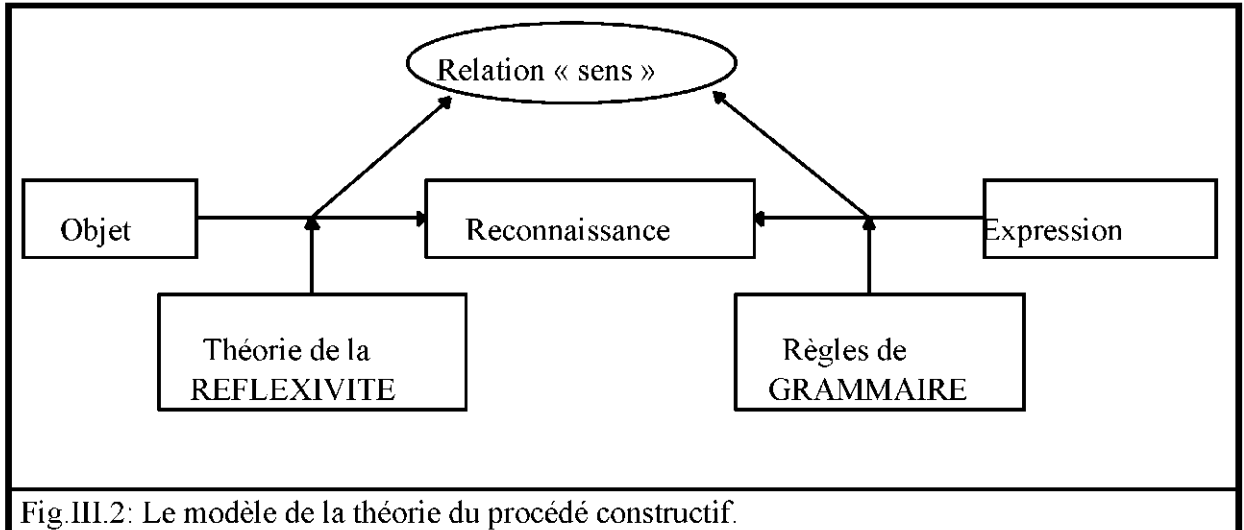
La Théorie du Procédé Constructif (TCP) est issue des travaux d'un grande école linguistique japonaise (cf. Pr M. Miyazaki, Japon IA N° 28).

Elle s'appuie sur la grammaire générale de Port-Royal (1660) et plus récemment, sur la grammaire générative de N. Chomsky (1950). Un des fondements essentiels de cette théorie réside dans la théorie du sens.

Selon cette théorie, trois processus interagissent:

- 1-Le monde des objets du discours,
- 2-la reconnaissance de ce monde par son interlocuteur,
- 3-leur expression (des interlocuteurs).

Une relation dynamique s'établit entre ces trois pôles pour construire un sens intelligible par des êtres en communication. (fig.III.2)



I.4 Mécanisme de traduction du système:

La TCP repose elle-même sur une méthode de traductions multiniveaux.

Les connaissances du traducteur en matière de catégories subjectives sont mises en relation avec le caractère subjectif de la phrase à traduire.

Ensuite, le caractère objectif de cette phrase est rendu objectivement en LC (exp. l'anglais), en respectant les niveaux idiomatiques, sémantiques de valence verbale et standard.

Enfin, les catégories subjectives servent à former une phrase anglaise correcte à partir de la structure objective de la phrase.

Remarques:

- 1-Pour la langue japonaise, langue agglutinante, l'expression finale englobe une partie objective représentée par les noms, les verbes et les

adjectifs, et une partie subjective représentée par des "joshis" , ie: des suffixes.

2-Le japonais accumule les mots et n'indique l'inflexion subjective qu'à l'aide de ces suffixes ajoutés au verbe; alors que les langues indo-européennes (anglais, français,...) sont construites d'une manière diamétralement différente et affectent aux signes linguistiques des catégories.

3-Diverses théories et méthodologies sont mis en œuvre pour le traitement de la langue japonaise.

4-La rapidité de traduction est observée avec les ordinateurs à structure parallèle ou massivement parallèle pour la résolution des ambiguïtés syntaxiques et la génération.

II. Le système ARIANE:

Le système ARIANE est développé par le groupe (ex-CETA) GETA à Grenoble en France en 1978.

Le système ARIANE 78.4 est destiné à développer des programmes de traduction de documents écrits dans diverses langues naturelles.

Il appartient à la famille des systèmes de deuxième génération utilisant la technique de transfert, avec beaucoup d'améliorations.

Le système ARIANE est un environnement complet de programmation pour la TAO, il comporte: Quatre langages spécialisés pour la programmation linguistique: TRANSF, ROBRA, ATEF, SYGMOR.

1-Un moniteur interactif est l'interface entre les utilisateurs et tout le système. Il assure le dialogue avec les utilisateurs en demandant des

paramètres et enchaîne les programmes du système en réalisant des tâches complexes en fonction des paramètres obtenus.

2-Un ensemble de programmes utilitaires qui facilitent la mise au point et l'exploitation des programmes linguistiques.

II.1 Le processus de traduction:

Le processus de traduction d'une LS vers une LC est divisé en 3 étapes logiques:

- 1-Analyse
- 2-Transfert
- 3- Génération.

Chacune étant décomposée en au moins deux phases successives, comme l'illustre le diagramme ci-dessous: fig.III.2.1 .

Phase logique	phase physique	langage (LSPL)
ANALYSE :	Analyse Morphologique : AM	ATEF
	Complement Lexical : AX	EXPANS
	Complement Lexical : AY	EXPANS
	Analyse Structural : AS	ROBRA
TRANSFERT :	Transfert Lexical : TL	EXPANS
	Complement Lexical : TX	EXPANS
	Transfert Lexical : TS	ROBRA
	Complement Lexical : TY	EXPANS
GENERATION :	Complement Lexical : GX	EXPANS
	Génération Structurale : GS	ROBRA
	Complement Lexical : GY	EXPANS
	Génération Morphologique : GM	SYGMOR

Fig.III.2.1: Les phases du système.

II.1.1 Analyse:

Le texte d'entrée en LS est soumis à l'Analyseur Morphologique (AM), écrit en ATEF, qui produit comme résultat un arbre décoré.

Ce résultat est ensuite transformé en un autre arbre (descripteur multiniveau) par l'Analyseur Structural (AS), écrit en ROBRA.

II-1-2-Transfert:

a) Transfert Lexical: TL

Le TL est écrit en TRANSF. Il transforme l'arbre au moyen d'une consultation des dictionnaires bilingues multichoix. Le traitement des polysémies et des tournures complexes est possible.

Transfert Structural: TS

Le TS traite les différences structurales entre 2 langues, en rendant la structure linguistique intermédiaire conforme à la nature de la langue cible.

Cette phase finit également une partie du TL (choix d'équivalents en fonction d'un contexte étendu, traitement des tournures par le TL), et traite les phénomènes qui n'ont pas pu être calculés assez profondément par l'AS.

II-1-3-Génération:

a) Le Générateur Syntaxique: GS

A partir du descripteur produit par le transfert, le GS calcule :

- Une structure de surface qui est une paraphrase possible du texte (contraintes par les impératifs de la traduction). Concrètement, elle recalcule les fonctions et les classes syntaxiques à partir des relations logiques et sémantiques.

- L'ordre des mots.

- La génération des mots « outils » absents et de toutes les questions liées à l'actualisation.

b) Le Générateur Morphologique: GM

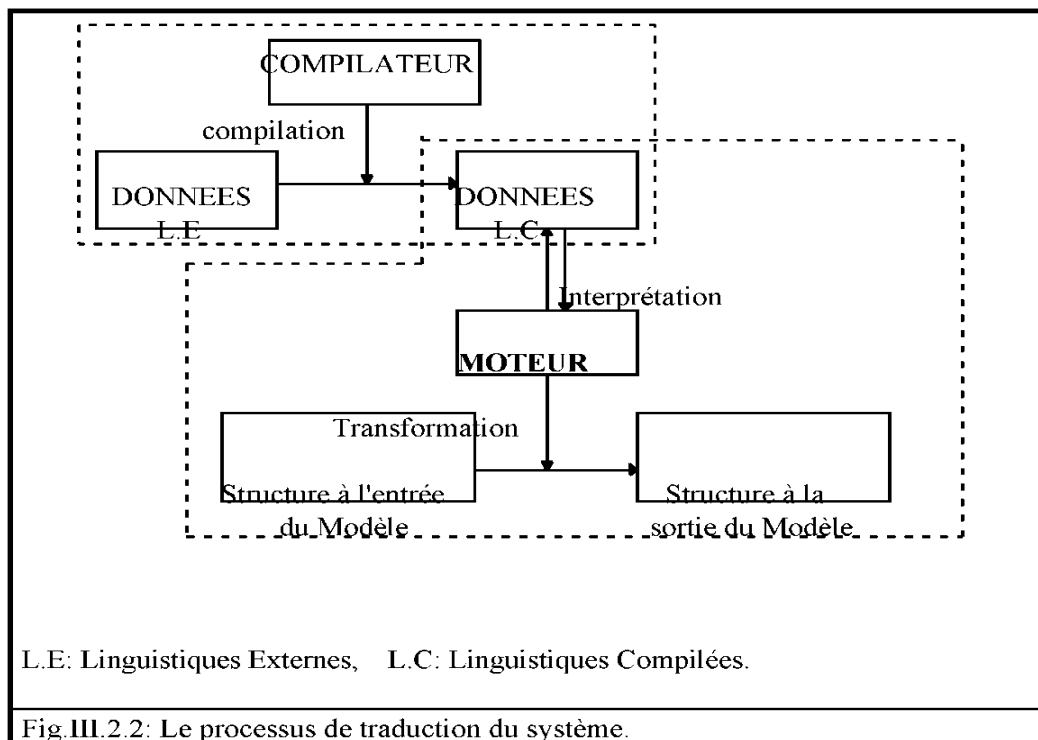
Il travaille sur les feuilles de l'arbre (ou terminaux), produit par le GS, et produit des formes morphologiques correctes constituant un texte en LC .

Chaque étape est écrite dans un LSPL réalisé à l'aide d'un compilateur et d'un moteur interprétant les tables produites par le compilateur.

II.2 Evolution d'ARIANE: ARIANE-G5, 1988

Depuis 1988, une nouvelle version d'ARIANE a été mise en service ARIANE-G5.

Le processus de traduction est toujours subdivisé en 3 étapes logiques; chacune des étapes est décomposée en au moins deux phases successives comme l'illustre le tableau fig.III.2.2 .



II.3 Utilisation de l'aspect multilingue d'ARIANE:

Une des particularités importantes du système ARIANE réside dans son utilisation multilingue.

Ainsi pour développer un système complet multilingue avec N langues, il faut réaliser:

- N analyses ,
- $N*(N-1)$ transferts,
- N générations .

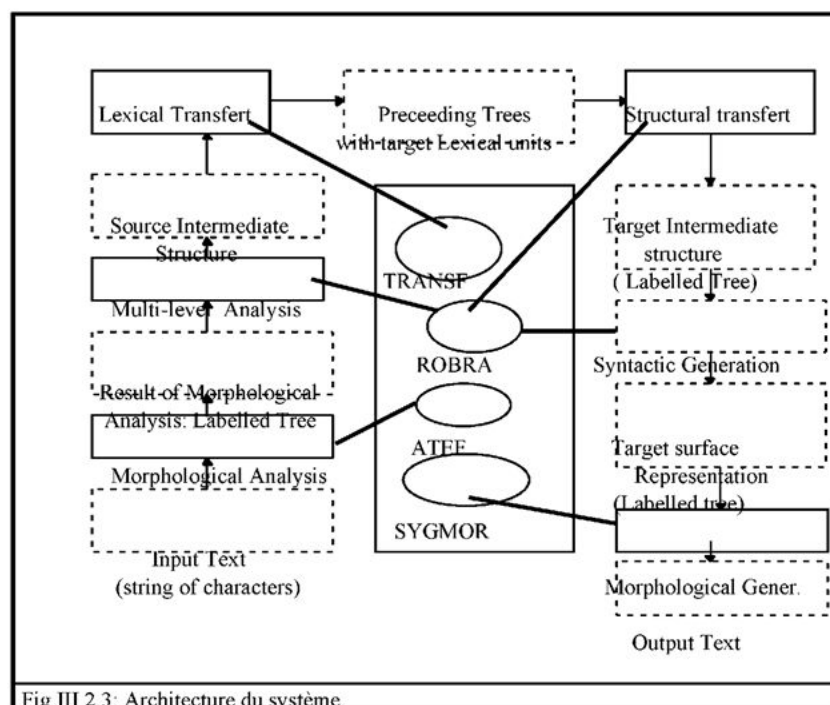
II.4 Application - traducteur de l'anglais vers l'arabe :

Grâce à l'aspect multilingue du système, Moneimne a réutilisé l'analyse de la LS: anglais pour la génération vers l'arabe (fig.III.2.3).

a) Développement du transfert vers l'arabe:

Dans l'approche du GETA, la phase doit être la plus courte possible. Elle se limite aux dictionnaires bilingues: TL et à un TS minimal. Cette particularité permet de gagner du temps dans le développement de nouveaux couples de langues.

L'architecture du système avec ses différentes phases : (fig.III.2.3).



b) Développement de la génération de l'arabe:

La génération de l'arabe est une des composantes les plus importantes du travail mené par Moneimne en 1989. L'utilisation de la langue arabe dans le système ARIANE représentait une dure épreuve pour:

- les tests sur le système lui-même,
- la théorie linguistique du GETA .

Toutes les possibilités du système sont utilisées pour développer la génération syntaxique (GS) .

Pour la GM, il a fallu faire un traitement particulier des verbes compte tenu du système dérivationnel de la langue arabe. Cette partie a été traitée par le langage ROBRA.

Enfin, une spécification de la grammaire arabe, sous forme de grammaire statique, a été élaborée .

Remarques:

1-L'abandon par le GETA du concept de langage pivot contenu dans le projet du

CETA. En fait, cette langue pivot est la chimère des projets de TA/TAO: une représentation machine unique et non ambiguë qu'on donnerait à tout texte pour le traduire en n'importe quelle langue. Une telle représentation nécessiterait une représentation totale du signifié, valeurs pragmatiques comprises.

2-La caractéristique essentielle du langage informatique de haut niveau ROBRA, utilisé en analyse, en transfert structural et en génération syntaxique, est sa capacité d'analyse récursive.

Cela permet au système, lorsqu'il s'aperçoit qu'il suit "*un mauvais cheminement*" à travers le graphe de contrôle des sous-grammaires de revenir en arrière et de défaire une partie du travail pour prendre un autre cheminement possible. Cette faculté risque naturellement d'engendrer un certain bouclage si aucune voie ne semble aboutir.

3-La maîtrise des procédures permettant d'interrompre certains traitements et de produire des traductions "*sous-optimales*" à différents niveaux, semble particulièrement délicate.

4- Sur le plan calcul, le système ARIANE réalise 1,5 million d'opérations pour traduire chaque mot dans le traitement.

III. Le système AlethTrad:

AlethTrad, qui est un système d'aide à la traduction, développé par GSI-Erli, a débuté aux USA et en France (1993).

AlethTrad est un système de traduction humaine assistée par la machine. Le processus de traduction est divisé en deux phases:

1-L'analyse des textes,

2-La traduction - révision - validation.

III.1.L'analyse des textes:

Cette étape a pour objet:

Le résultat de cette traduction est un fichier qui présente les phrases du texte source suivies de leur traduction.

III.3.3. Traduction multi-niveaux:

Après , la phase du traducteur/réviseur et pour apporter d'éventuelles corrections, le texte est livré au client, et au stockage en mémoire de traduction. Voir fig.III.3.1.

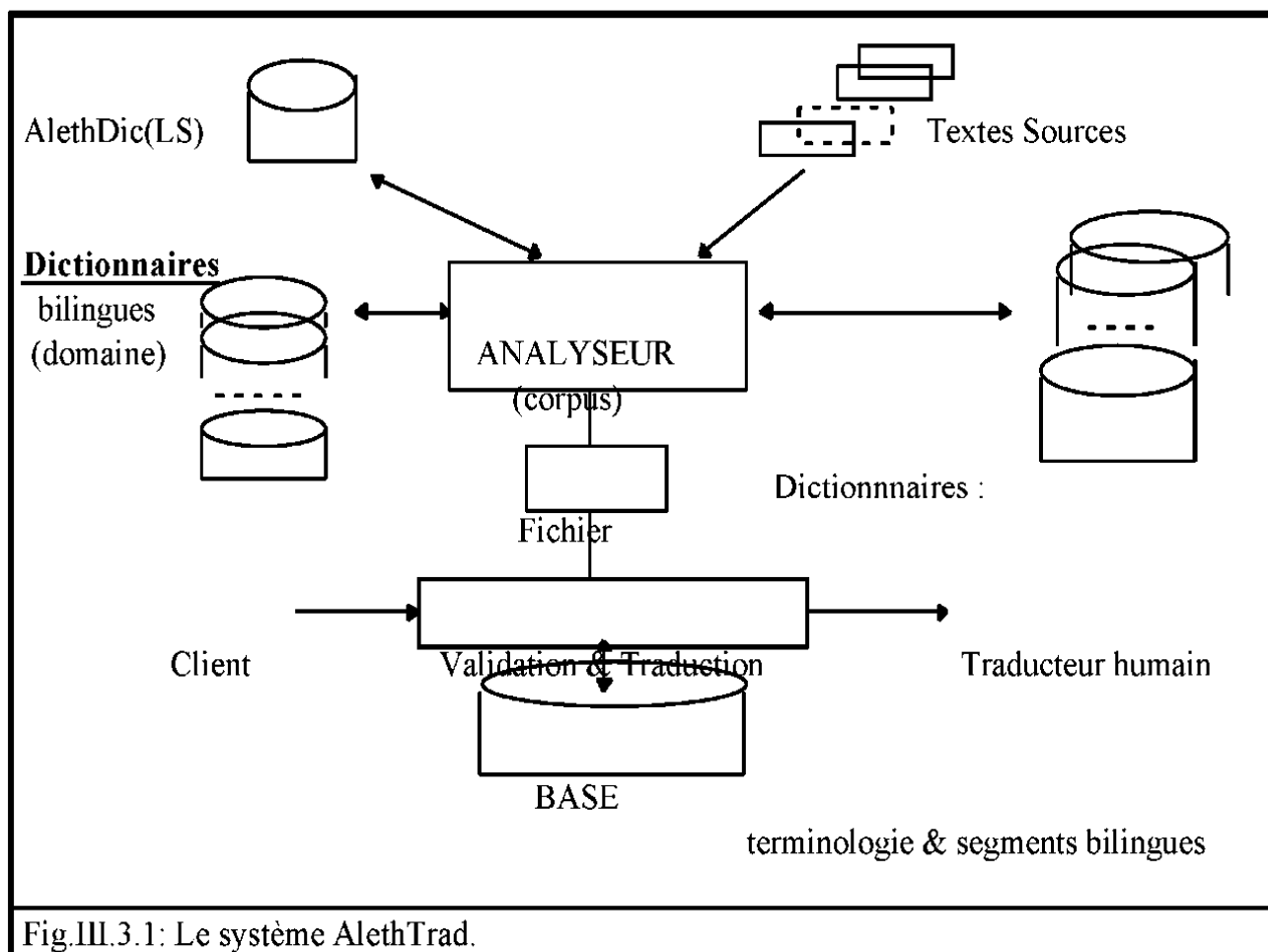


Fig.III.3.1: Le système AlethTrad.

Au niveau de développement actuel, AlethTrad travaille sur les trois premiers niveaux:

a) Niveau 1: Le traitement

- Chercher dans une mémoire de traduction une phrase correspondant à celle en cours d'analyse.

- Dans le cas où aucun résultat n'est obtenu, le système déclenche un algorithme d'appariement approximatif (fuzzy matching) qui va lui permettre de faire une recherche sur des chaînes peu différentes (exp. les numériques, mots de types grammaticaux, variantes flexionnelles...).

Ainsi, des séquences de phrases à traduire déjà présentes dans la mémoire de traduction seront traitées automatiquement.

b) Niveau 2: *analyse morphologique*

Si le premier niveau échoue, le système passe au traitement du deuxième niveau qui effectue une analyse morphologique sur des éléments de la phrase .

Il procède à une lemmatisation :

- les noms : leur forme au singulier,
- les adjectifs : masculin | féminin ,
- les verbes : l'infinitif, etc.

Puis le système construit des groupes de mots pour tenter un traitement du niveau-1.

Le résultat de l'analyse prend la forme d'une succession d'éléments phraséologiques ou bien d'éléments phraséologiques suivis d'éléments simples.

Dans le premier cas, le système déclenche une traduction directe à partir de la juxtaposition d'éléments phraséologiques. Dans le deuxième cas, le système déclenche un nouvel examen .

Niveau 3: *Le moteur de traduction*

Ce niveau intervient sur le résidu du document non-traduit par l'approche « *mémoire de traduction* » .

Le moteur déclenche des algorithmes impliquant l'utilisation des grammaires d'analyse, de transfert et de génération. La traduction se déclenche en 4 étapes :

1-L'analyse morphologique : .

2-L'analyse syntaxique : Le travail, à ce niveau, consiste à identifier certains groupes ou syntagmes (SN,SP,SV) sur lesquels interviendront des transformations structurelles.

3-Les transformations structurelles : Une fois les structures identifiées, celles-ci sont transformées en structures de la langue cible.

Pour ce faire, le système s'appuie sur des règles de transformation guidées par des informations contenues dans les dictionnaires multilingues.

Les règles de traduction implémentées traitent des cas génériques. Cependant, elles cherchent à couvrir les phénomènes linguistiques identifiés dans les documents à traduire .

4-La génération : Elle est prise en charge par un automate qui exploite les informations de ces structures.

Le système opère des substitutions lexicales, des insertions d'éléments lexicaux, générer les flexions des mots en langue cible , etc.

III.4 Architecture d'AlethTrad: (Voir *fig.iii.3.2*)

AlethTrad a été développé en langage C et il tourne sur des machines Unix (stations de travail). Il travaille en mode serveur: des postes de traduction en réseau (Unix, Mac, PC) .

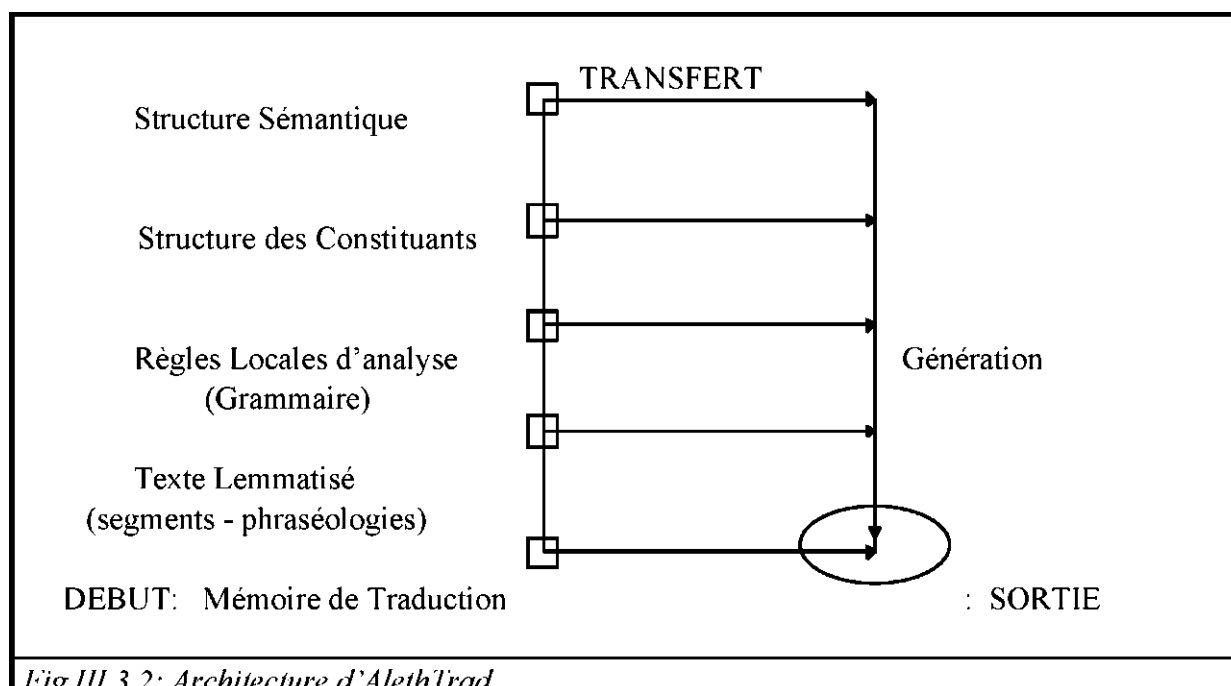


Fig.III.3.2: Architecture d'AlethTrad.

Le logiciel d'édition de textes utilisé en interface par les traducteurs est FRAMBuilder s'offre des fonctionnalités supplémentaires, comme la gestion des documents structurés en format SGML ou la communication de ces documents à d'autres logiciels par intermédiaire d'un API .

Une interface Word est en cours de développement et le premier bilan d'AlethTrad est de plus 40.000 pages traduites.

Remarques:

- 1-AlethTrad n'est pas un système de TA/TAO, mais un Outil d'Aide à la Traduction.
- 2-Le système est adapté aux besoins des utilisateurs.
- 3-La phase de post-traitement consiste à adapter le texte traduit vis-à-vis du texte source en recherchant, par exemple, des calques de structures et d'attributs. Ce point constitue une préoccupation.

4-La génération en langue cible est prise en charge par un automate, en opérant par des substitutions (lexicales, insertions d'éléments lexicaux, génération des flexions,...). Chacune de ces opérations est en cours d'innovations.

IV. Le projet METAL:

Origine: Université de Texas - USA et SIEMENS en Allemagne . *Période:* 1961-1990 .

IV.1 Historique et évolution de METAL:

Année	Langage de programmation & Machine
1961 - 1970	Fortran, CDC 6600
mi- 1970	Lisp UCI-Lisp InterLisp, DEC-10 InterLisp, DEC System-20
fin- 1980	InterLisp, Symbolics Computers LM-2

IV.2 Le système METAL:

Le système METAL traduit un texte d'une LS vers une LC par analyse automatique :

1-Morphologique,

2-Syntaxique,

3-Sémantico-syntaxique : relations et liens exprimés en LS.

METAL est qualifié de système intermédiaire de 2-ème et 3-ème génération.

a) Les outils du développement de la grammaire :

La composante linguistique qu'utilise METAL est une grammaire à structure de phrase augmentée, classiquement c'est la forme de description selon Winograd, 1983.

Dans cette grammaire, les règles s'appliquent lorsque la structure décrite est rencontrée et que certaines conditions sur les constituants de la structure sont vérifiées par l'occurrence des fonctions des constituants.

S'il y a succès, la règle construit un nœud gauche de l'arbre, qui a certaines caractéristiques spécifiques, généralement des Traits de Valeur (TdV).

Une des actions de base dans la construction de l'arbre est le passage relevant des TdV du nœud fils au nœud père et le transfert père-fils.

b) Traduction de phrases :

Dans ce cas, la phrase identifiée par le système permet la production de plusieurs interprétations et donc leurs traductions . Le système procède par étapes successives :

© : L'appel de la fonction TRANSLATE (appel à l'environ. Lisp).

© : La mise sous forme d'une liste de la phrase.

© : Le nombre d'interprétations de l'analyseur.

© : Le transfert et la génération .

© : La traduction de l'interprétation préférée.

Fig.III.4.1: Les étapes du système METAL.

On illustre par un exemple la traduction de l'allemand vers l'anglais:

© (translate)

© Sentence: (Ein alter, verheirater Mann hat dieses Programm niemals geschrieben.)

© 1 interpretation in 25314 ms. 286 phrases: 250 rejected.

© Transfert plus generation time: 14538 ms.

© (|an| |old| |married| |man| |has| |never| |written| |this| |program| .)

Fig.III.4.2: Traduction de l'allemand vers l'anglais par METAL.

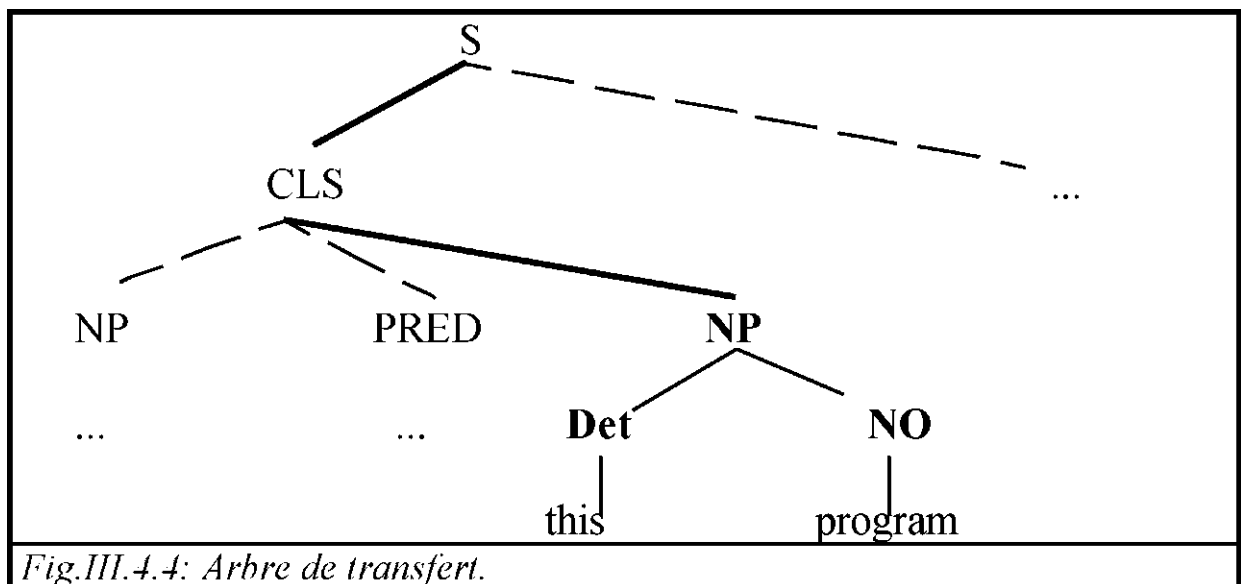
c) Le « P-TREE » :

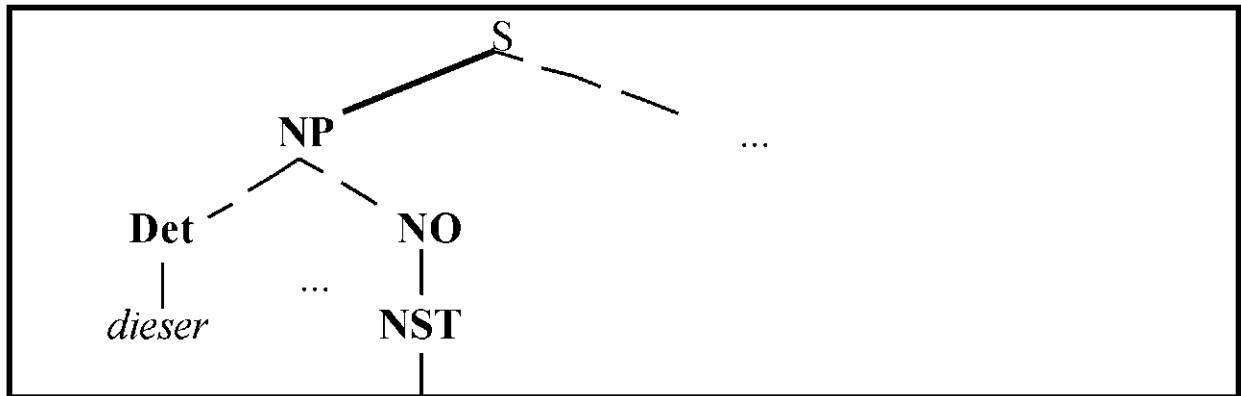
Le P-tree ou P-arbre ou arbre décoré est la forme d'interprétation étalée communément utilisée par les développeurs du système. Il apporte du sens en visualisant la phrase analysée.

Le P-tree prend obligatoirement les arguments de base et ceux optionnels. Le premier argument (de base) est le nombre des interprétations, le second est le « tree-type » .

Le tree-type étale la forme de la structure de la phrase après la construction de tous les nœuds et l'application des règles sur la structure .

Le P-tree peut être visualisé soit en phase d'analyse soit en phase de transfert. Voir [Fig.III.4.3](#) et [Fig.III.4.4](#) .





IV.3 Les outils du développement lexical:

a) L'inter-codeur :

L'interface entre le système METAL et l'utilisateur est le moyen de coder et de modifier le lexique.

Quand l'inter-codeur est appelé pour consulter un mot ou un terme en LS et si ce mot n'existait pas dans le lexique du système, celui-ci crée une entrée dans la base, en présentant un menu complet pour remplir les différents champs correspondant et c'est ainsi que l'utilisateur peut intervenir .

b) Les fonctions lexicales :

Le système METAL permet l'accès à la BD lexicale dans un mode différent de celui d'un utilisateur averti . Cet accès demande des moyens d'aides et en quantité de lexique.

Ainsi, la création et l'installation d'un ensemble terminologique ou le recodage de la base existent quand il y a changement ou ajout dans la grammaire du système. Ceci demande une forme d'accès qui importe des mesures de sûreté et une exécution en batch.

Les fonctions lexicales sont converties par le détecteur d'erreur dans la base « Defaulter », le pré-analyseur « Preanalyser » et le valideur « Validator » pour le lexique en batch.

IV.4 Evolutions:

Les recherches associées au projet METAL ont pris en considération le fait que le système s'oriente à des fins de traductions industrielles.

Pour ce faire, la concentration des travaux menés par les linguistes et les informaticiens assure l'efficacité des résultats.

Les outils de travail au sein du projet sont implémentés dans un langage structuré et ayant un format familier pour les linguistes.

Ainsi, l'ajout, la modification et la suppression de certaines structures sont facilement manipulables à l'expérimentation et les structures deviennent de plus en plus efficaces.

Ces outils restent utiles pour la création d'un nouveau langage et d'un nouveau système de traitement automatique de la langue (NLP-system) possible et efficient.

Remarques:

- 1-Métal est un système de traduction assistée par ordinateur, le processus de traduction comporte une phase de post-édition.
- 2-Le même résultat de l'analyse, pour une langue donnée, est utilisé pour la génération dans plusieurs autres langues: le concept d'analyse permet d'effectuer des traductions multilingues.
- 3-Enrichissement permanent des dictionnaires.

4-Métal est initialement conçu pour traduire de l'allemand vers l'anglais, puis étendu vers d'autres langues européennes.

Les résultats diffusés ne montrent pas la capacité du système pour la traduction dans des langues de classes différentes, comme l'arabe.

V. Le système TOSHIBA :

V.1 Introduction:

Le système TOSHIBA de TA fonctionne sur un mini-ordinateur 32-bits (UX-700) , et traduit des textes scientifiques et techniques de l'anglais vers le japonais.

Il est basé sur la technique de transfert sémantique, pour ceci un modèle de grammaire appelé LTNG:« Lexical Transition Network Grammar » a été développé.

Le système comporte 3 types de dictionnaires:

- 1-Un dictionnaire général ,
- 2-Un dictionnaire pour les termes techniques,
- 3-Un dictionnaire utilisateur.

Le système est équipé d'un éditeur bilingue anglais/japonais pour le pré/post-édition et des utilitaires (logiciels).

V.2 Les caractéristiques du système:

Le système comporte :

- Un puissant analyseur sémantique,
- un moteur de traduction (Parser): le parseur,
- un éditeur bilingue.

L'analyseur sémantique représente la signification de la phrase comme une structure conceptuelle de celle de la structure syntaxique et de la signification des mots qui apparaissent dans la phrase.

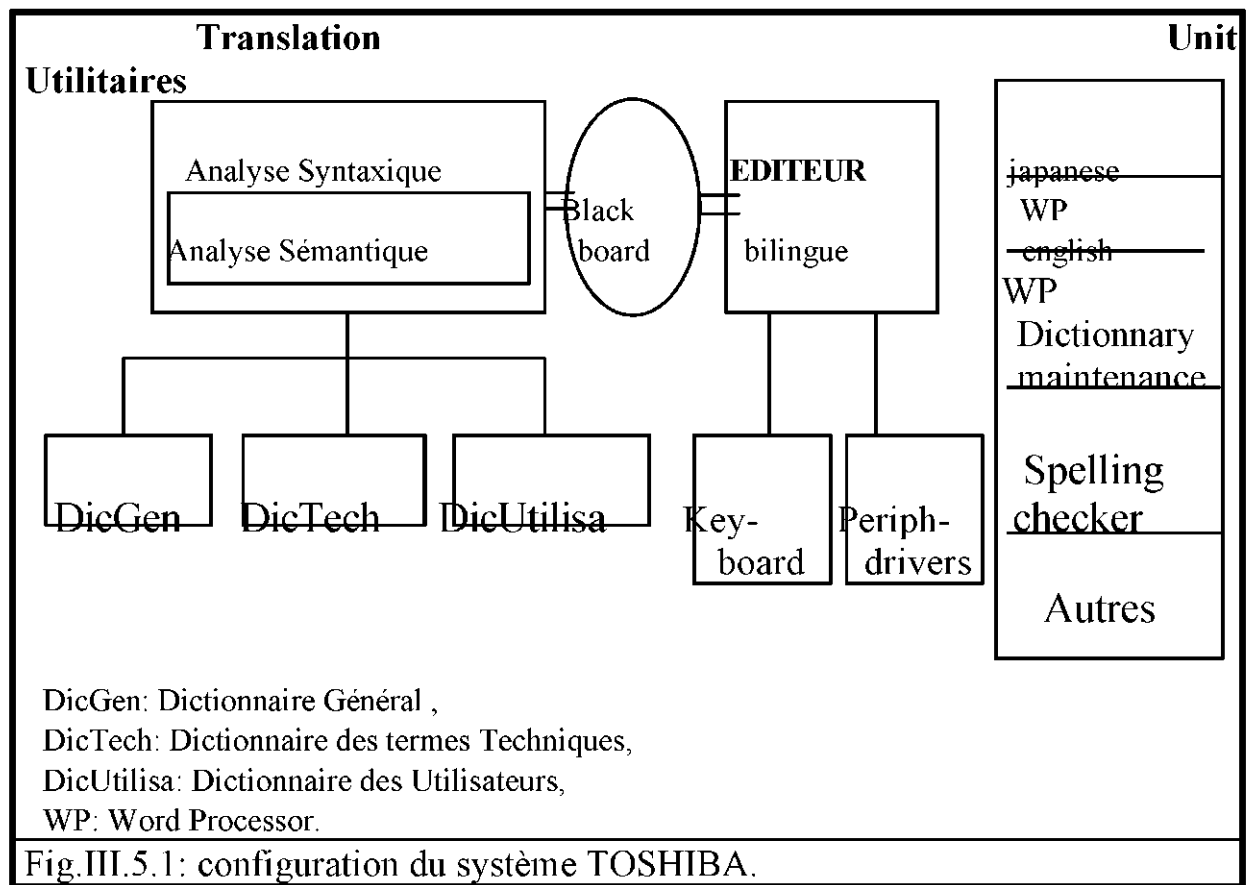
Le travail du système est en conjonction avec l'analyseur sémantique et le parseur avec une stricte séparation entre analyseur syntaxique et sémantique

.

L'éditeur bilingue est une interface homme-machine avancée, entre l'opérateur humain et le système TOSHIBA.

V.3 Configuration du système:

Le système de TOSHIBA est écrit en langage C et sous Unix (UX-700) et a comme configuration celle du schéma de la Fig.III.5.1 .



V.4 La méthode de traduction :

V.4.1 Analyse morphologique - dictionnaires :

Il y a 3 types essentiels de dictionnaires utilisés par le système:

- 1-Le dictionnaire général: il comporte environ 50.000 mots d'usage général et d'idiomes.

2-Le dictionnaire des termes techniques: il comporte au maximum 50.000 termes techniques partitionnés selon des thèmes : électronique, mécanique...

3-Le dictionnaire utilisateur: il comporte des mots spécifiques des utilisateurs qui seront sauvegardés et éventuellement leur traduction.

Les mots dans une phrase anglaise apparaissent avec une morphologie complexe. L'analyseur morphologique divise le mot en morphèmes et construit une structure de mots . (fig.III.5.2)

1 Reconnaissance des inflexions morphémiques dans un mot :

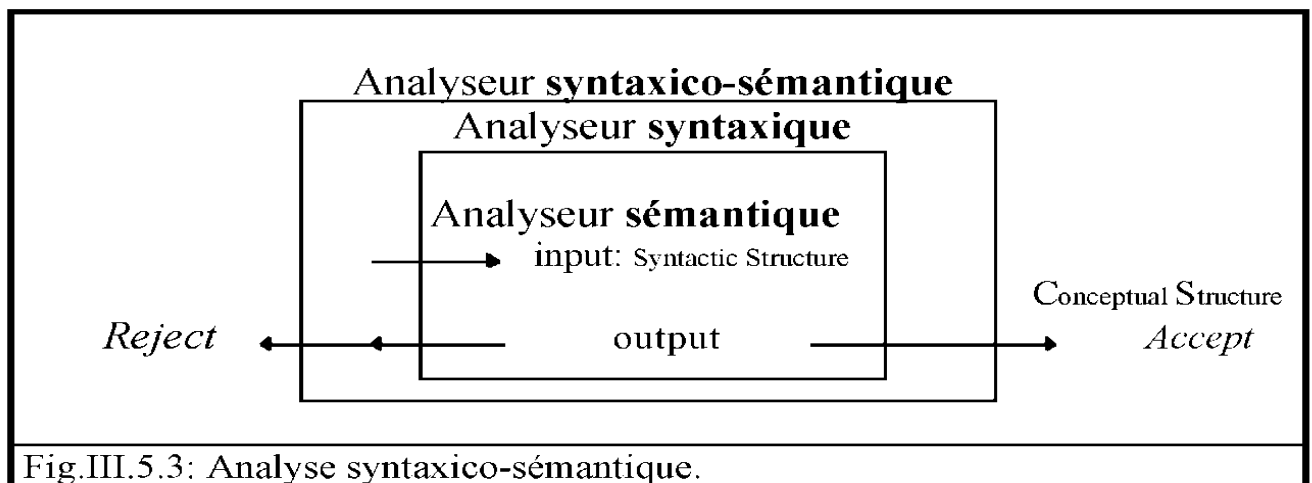
exp. comme le pluriel '(e)s' et le passé '(e)d' en anglais

2 Construction de la structure de mots.

| Fig.III.5.2: L'analyseur morphologique

V.4.2 Analyse Syntaxique: ASy

Les analyses syntaxique et sémantique sont séparées et non complètement



Les traits de l'analyse syntaxique du système

indépendantes : elles fonctionnent séquentiellement et procèdent dans un mode interactif. (Fig.III.5.3)

1- Une seule structure syntaxique est dérivée de l'ASy. L'ambiguïté est résolue en éliminant les incohérences des catégories des mots et leurs combinaisons.

Si l'ambiguïté persiste, l'analyse sémantique construira une structure conceptuelle plausible et résoudra implicitement les ambiguïtés.

2- L'analyse syntaxique est purement syntaxique et les règles syntaxiques ont des conditions sémantiques .

3- L'analyse syntaxique est composée de deux parties :

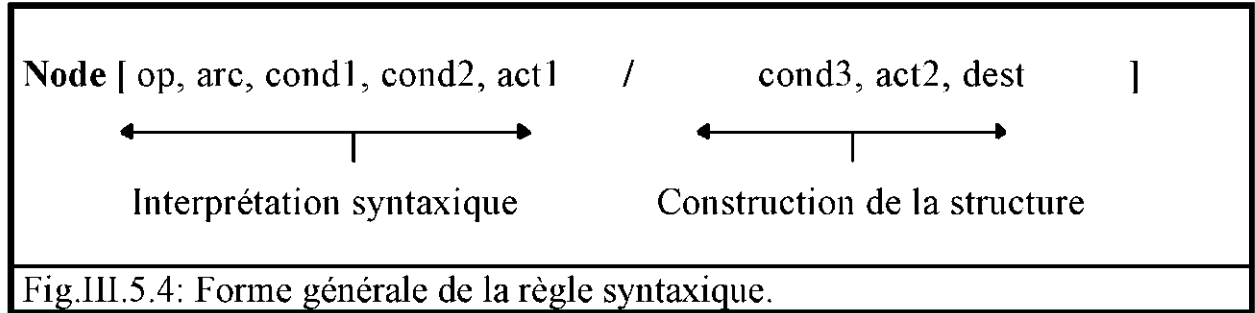
- l'Interprétation Syntaxique : IS ,
- la Construction de la Structure Syntaxique .

L'interprétation syntaxique des phrases en entrée est séparée du processus de construction de la structure .

L'IS consiste à rechercher les règles syntaxiques à employer et de cette façon les ambiguïtés lexicales sont résolues.

La construction de la structure applique ces règles à la manière d'un parseur pour construire la représentation syntaxique .

Les règles syntaxiques sont décrites par une grammaire ATN (Augmented Transition Network). La fig.III.5.4 , montre la forme générale d'une règle.



V.4.3 Analyse Sémantique: ASé

L'analyse sémantique construit une interprétation sémantique et des structures conceptuelles pour la langue cible.

Si une "Interlangue" est choisie comme langue cible alors le système de traduction devient un système à interlangue.

La méthode de l'analyse sémantique est basée sur les hypothèses suivantes :

- HYP1: "le sens est lexical".
- STANDPOINT1: Syntaxe "les règles relatives à l'ordre des mots".
- STANDPOINT2: Syntaxe "les règles relatives à l'ordre des mots et une partie de la sémantique".

Notions sur les règles sémantiques :

Les règles sémantiques consistent en une conversion arbre-à-arbre (tree to tree conversion) avec des conditions et actions exécutées .

La forme d'une règle sémantique est comme suit:

$$MP = TP ; COND ; ACT ; CTRL$$

où :

MP « Matching Pattern »: correspondant à l'arbre de la structure syntaxique. TP « Target Pattern »: correspondant à l'arbre à générer pour la cible.

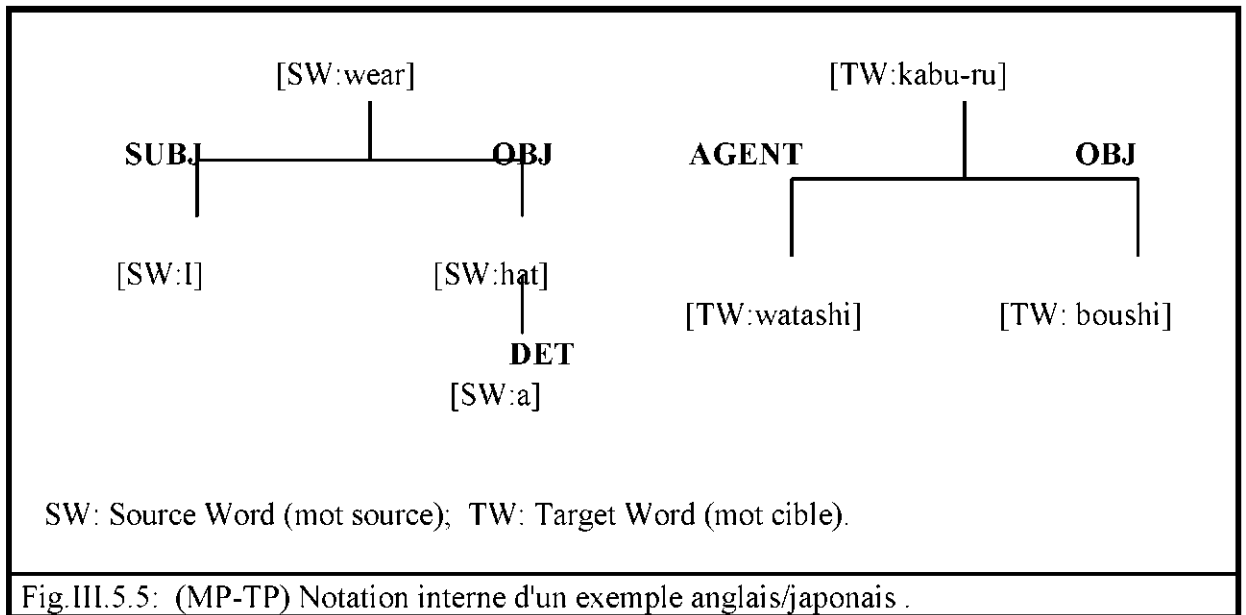
COND « CONDitions »: conditions de conversion de MP à TP. ACT « ACTions »: actions exécutées pour la conversion.

CTRL « ConTRoL »: contrôle du flux des règles lexicales attachées à chaque mot.

MP et TP représentent un arbre, utilisant la même notion pour un arbre syntaxique , i.e. les nœuds sont connectés par des arcs nommés et le nom de l'arc dénote une relation. Voir fig.III.5.5

V.4.4 Transfert structural:

Le transfert lexical est évoqué par les mots ou termes



Le transfert structural est un autre type de transfert qui est évoqué par les traits des structures syntaxiques .

Le transfert lexical est une sorte de conversion de la syntaxe à la sémantique alors que le transfert structural est une conversion de la langue source à la cible et par conséquent une réalisation de grammaire contrastive .

Un exemple de transfert structural est le traitement de la différence dans le mode "temps" entre le japonais et l'anglais .

L'assignement correct du temps est traité dans le transfert structural qui est syntaxique, non sémantique et dépend d'une grammaire contrastive.

V.4.5 La génération:

V.4.5.1 La génération syntaxique:

Le rôle de la génération syntaxique est de :

- 1-Déterminer l'ordre des mots.
- 2-Attacher des postpositions (Joshi en japonais).

Les règles de génération décrivent ces deux fonctions. La procédure de génération commence par la règle attachée à la phrase et le noeud tête de la structure conceptuelle.

V.4.5.2 La génération morphologique:

En général, la génération morphologique est inscrite dans les dictionnaires .

La façon simple d'enregistrer la traduction des mots est de les mettre dans le champ TW: Target Word de la structure des mots. Cette méthode est employée quand la traduction équivalente est non-inflexionnelle, comme les noms.

Dans le cas de la traduction inflexionnelle, comme les verbes ,les adjectifs... , l'information inflexionnelle doit être présente.

On illustre par l'exemple suivant :

la traduction du mot « read » est « yo-mu », et les informations utilisées pour cette génération sont:

© stem : racine « yo »

© conjugation type : «5-dan»

(D kind of conjugational part : « ma »

© other information (irregular form) : « onbin-kei » .

Le système emploie six catégories morphologiques pour le japonais: (tense) temps, aspect, (modality) modalité, (factitive) factif, passive et négation .

L'exemple, fig.III.5.6, illustre une génération morphologique.

INPUT : « I could not run . »

Morphological analysis: [SW: I] [SW: can, TENSE: past] [SW: not]
[SW: run, TENSE: present]

Syntactic analysis : [SW: I]
[SW: run, MODALITY: (can (TENSE:past))
TYPE : négation]

Fig.III.5.6: Informations pour la génération morphologique.

Remarques:

1-La langue japonaise est très différentes des langues indo-européennes, notamment par sa morphologie externe.

2-Le besoin en systèmes de TA/TAO au japon est important, ceci pour les échanges technologiques et industriels avec les autres pays du monde.

3-Le domaine de recherche en linguistique et TALN sont largement considérés, particulièrement au développement de la morphologie (les dictionnaires électroniques et les bases de données morphologiques) et la sémantique (les grammaires appropriées au traitement de la sémantique).

VI. Le système TORJOMAN:

TORJOMAN est un système de traduction assistée par ordinateur Anglais/Arabe.

Origine : IRSIT, Institut de Recherche des Sciences Informatiques et des Télécommunications en Tunisie. **Période:** 1989- 1995.

VI.1.Introduction:

Le système TORJOMAN de TAO fonctionne sur machine compatible IBM-PC sous DOS/Windows .

Il a pour but de traduire des structures de base de phrases anglaises et sans restriction à un domaine particulier.

Le système a débuté en 1989 à l'IRSIT en Tunisie avec la collaboration d'une firme américaine PC-Linguistics au Texas (USA).

VI-2-Aspects fonctionnels :

TORJOMAN est une première version d'un moteur de traduction automatique. C'est un système complet d'aide à la traduction dans le sens où il est capable de répondre à une demande de traduction, soit:

- 1-en donnant directement un équivalent arabe de la phrase anglaise,
- 2-en proposant à l'utilisateur des choix dans le vocabulaire de base,
- 3-en fournissant en dernier recours, un accès direct au dictionnaire des mots, pour une traduction mots à mots.

Comme le système TORJOMAN est un système de traduction assistée par ordinateur, ce dernier fait appel à une traduction extérieure :

- si le moteur de traduction se trouve confronté à plusieurs traductions possibles d'un termes anglais, il propose une liste de synonymes permettant à l'utilisateur un choix adéquat ,
- convivialité du système en donnant moyen à l'interprète humain d'affiner la traduction fournie par le système.

VI.3. Aspects techniques :

Le processus de traduction est articulé essentiellement autour de trois phases:

- 1-compilation des dictionnaires ,
- 2-analyse de l'anglais ,
- 3-transfert et génération .

VI.3.1. La compilation des dictionnaires :

Le moteur de traduction qui est à la base de TORJOMAN fonctionne grâce à trois types de dictionnaires :

- le dictionnaire général des mots ,
- le dictionnaire des expressions ,
- le dictionnaire de l'utilisateur.

■ Dans le cas où la phrase à traduire correspond à une expression idiomatique, TORJOMAN opère automatiquement sa traduction à partir du **dictionnaire des expressions**.

Exemple:

« like father like son » : *?iNNa HaDaA ?aL-SSiBL MiNDaAKa ?aL-?aSaD* .

Le système, dans ce cas, sait qu'il ne s'agit pas d'analyser la phrase car une expression idiomatique porte un sens global indépendant des termes qui la composent.

« Dans le cas où il ne s'agit pas d'une expression idiomatique entière, c'est le cas de traitement des expressions idiomatiques à usage verbal. En effet, le verbe en anglais peut être suivi d'une séquence spécifique de mots donnant un sens global au verbe différent de son sens en tant qu'unité autonome.

Torjoman opère par la détection de cette séquence puis sa traduction en l'intégrant au sens global de la phrase.

Exemple 1:

« the man **breaks a journey** » , si le système ne tient pas compte de l'usage idiomatique du verbe : *KaSSaRa ?aL-RaJuLu ?aL-RiHLa&* , alors qu'en réalité sa traduction est plutôt : *TaWaQQaFa ?aL-RaJuLu ?aTNaA?a ?aL-RiHLa&* . Ainsi, Torjoman reconnaît ce genre de tournure verbale.

Exemple 2:

« pull out »:
?iQTaLa^Ca . « call
off » : *?aLGaf*.

- Dans le cas où il ne s'agit pas d'une expression idiomatique ou idiomatique verbale, le système soumet la phrase à l'analyseur syntaxique pour ensuite effectuer les phases de transfert Anglais/Arabe et la génération en arabe (consultation du dico. général des mots).

« Quand la phrase contient un nom propre ou un acronyme, le système cherche sa traduction dans un dictionnaire spécial réservé à l'utilisateur. Ce dernier y introduit son vocabulaire propre et spécifique (**dico. de l'utilisateur**).

Exemple:

« IRSIT »: *Ma^CHaD ?aL-BuHuT Li-^CuLuM ?aL-
?i^CLaAMiYa& Wa ?aL-?iTtiSaALaAT^CaN Bu^CD.*

« En dernier recours, quand la phrase n'est pas acceptée par l'analyseur syntaxique, une traduction mot à mot est fournie.

VI.3.2. L'analyse de l'anglais :

Des algorithmes d'analyse et de génération forment le noyau du moteur de traduction et permettent au système de réaliser des tâches 'intelligentes' .

Le principal problème en analyse est celui de l'ambiguïté du langage naturel. Ces ambiguïtés peuvent prendre de formes variées:

- 1-lexicales : homographie, homonymie, polysémie.
- 2-structurelles : anaphore, coordination, rattachement.
- 3-sémantiques : sens , texte.

L'analyseur syntaxique de l'anglais est un automate ATN (W.Wood,1970 et T.Winograd,1983) sans le retour-arrière (backtracking), mais complété par

des règles ou heuristiques d'élimination des ambiguïtés lexicales (par filtrage et sélection au voisinage des mots).

Ainsi, la grammaire d'analyse est une grammaire transformationnelle qui construit l'arbre de dérivation en donnant la structure profonde de la phrase.

VI.3.3. Le transfert et la génération :

La stratégie de traduction utilisée par le système est la méthode de transfert, qui est fondée sur une description structurale des langues concernées (anglais/arabe). Cette phase de processus de traduction consiste à exploiter les résultats de l'analyse du langage source et à la définition des différentes structures grammaticales du langage cible puis la génération en langue cible.

VI.4. Caractéristiques du système TORJOMAN :

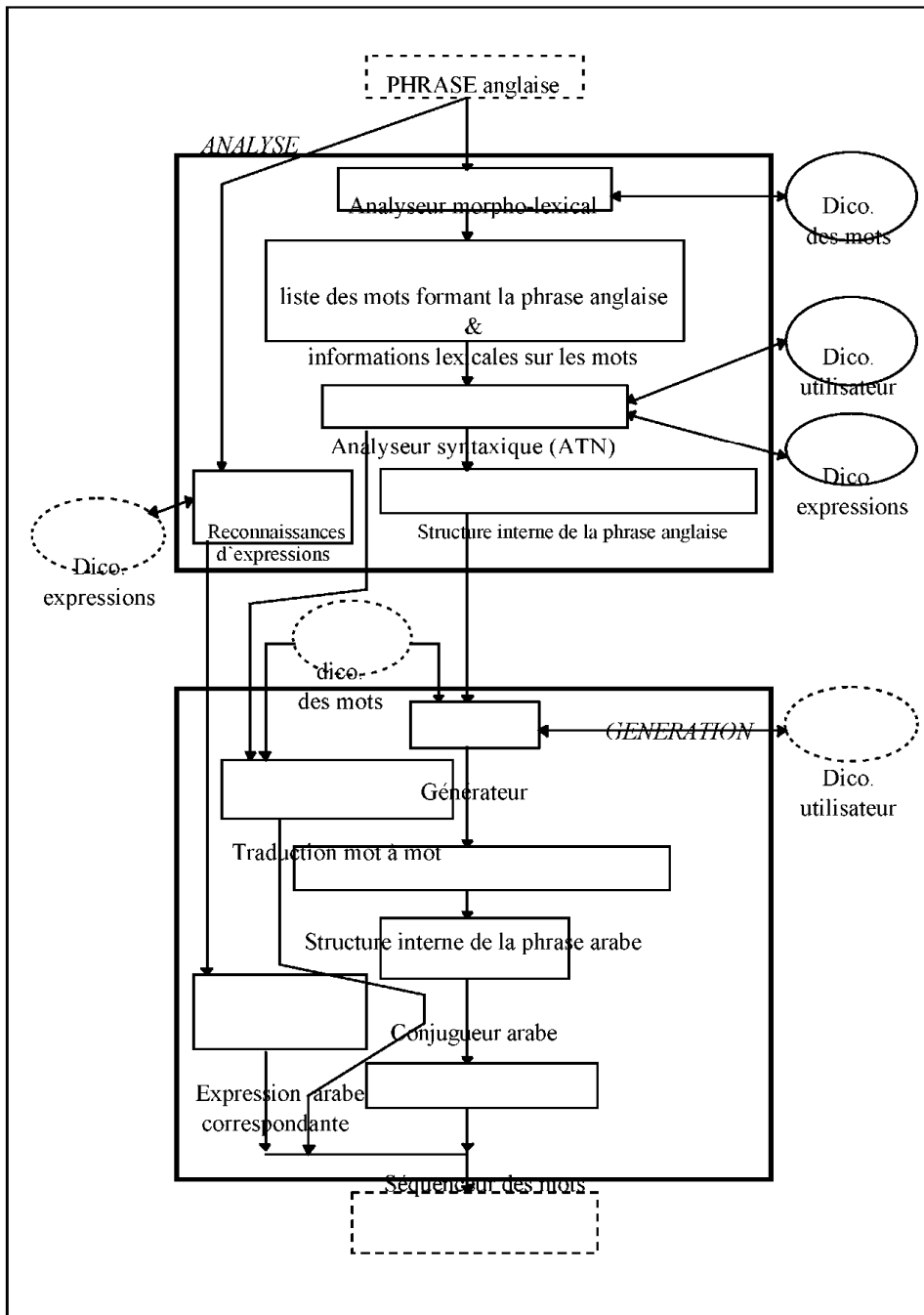
Torjoman dans des versions antérieures traduisait des phrases simples au mode déclaratif et la voix affirmative. Plusieurs améliorations ont permis l'évolution du système :

- Traitement des coordinations (and, or, but) ainsi la structure de la phrase traitée devient complexe et comporte plus qu'un verbe, adjectif, adverbe ... et plus qu'une phrase simple. Ce qui rapproche progressivement à la structure de texte.

L'introduction des coordinations au niveau des syntagmes a permis au système de tenir compte des cas de duel, tout en considérant les accords des adjectifs et des verbes y afférents.

- Traitement des noms composés complexes qui constituent une particularité de la langue anglaise, des subordinations et des expressions idiomatiques à usage non-verbal.

VI.5. Architecture du 'moteur' de traduction :



PHRASE ARABE

VI.6. Remarques sur le système TORJOMAN :

1-Le modèle théorique des grammaires transformationnelles, en l'implémentant sous forme des ATN, se heurtait à des problèmes d'implémentation :

Le résultat des recherches de S. Peters et R. Ritchie, 1973 ([ABEILLE 94]-p13) montrait que dans le cas général, on se heurtait à des problèmes d'indécidabilité si l'on tentait d'analyser une phrase incorrecte.

2-Les ATN eux-mêmes se révèlent mal adaptés :

Ils étaient lourds à programmer et à modifier, ils appartenaient à un style de programmation dit « procédural » où l'on décrit les objets par les procédures à utiliser pour les construire, mêlant connaissances linguistiques et instructions informatiques proprement dites.

3-Dans l'optique d'une analyse plus complète, le lien était difficile à faire avec le calcul d'une représentation sémantique.

4-Le dictionnaire général des mots ne permet pas l'accès direct à certaines dérivations d'un mot, comme dans le cas de traduction des noms composés complexes,...

VII-. le système SYSTRAN

VII-1.LE SYSTÈME

SYSTRAN est le premier système de traduction automatique. Dans les années 1965 /1985, il est passé par une étape qui a vu la dispersion de ses moyens et de ses droits. Chaque couple de langues constituait un bloc indissociable : pour traduire entre elles les 9 langues de la Communauté Européenne, il était nécessaire de développer 72 systèmes.

En 1986, la Société GACHOT a racheté les Sociétés créatrices WTC et LATSEC aux États-Unis, ainsi que SYSTRAN INSTITUT en Allemagne et SYSTRAN LUXEMBOURG. En regroupant cet ensemble sous la coupe de sa filiale SYSTRAN S.A., elle a formé une équipe unique qui a permis de restructurer le système sur de nouvelles bases. Celui-ci a retrouvé des possibilités de développement qui s'étaient perdues dans la diversité des besoins de ses clients multiples.

Les programmes sont découpés en trois ensembles : Analyse, Transfert et Synthèse. Chacun regroupe des modules indépendants réalisant une tâche bien déterminée.

1) L'analyse

C'est la partie la plus importante du programme qui représente pour un couple de langues pratiquement les deux tiers des 130000 lignes de programmation.

Son objectif est d'identifier la structure de la phrase en langue source, en

traitant lors de passes successives les divers types possibles d'interrelations syntaxiques les uns après les autres.

À chaque étape, elle prend des décisions sur l'existence ou la non-existence de ces interrelations entre les différents mots :

- levée des ambiguïtés homographiques,
- découpage des phrases en proposition principale et subordonnée,
- détermination des relations syntagmatiques de base,
- établissement des énumérations,
- détermination du prédicat principal,
- recherche du sujet et des antécédences,
- analyse des relations entre des propositions,
- recherche du gouvernement des prépositions.

2) La synthèse

Cet ensemble est chargé de restituer, selon le génie de la langue cible, ce qui a été analysé dans la langue source ; il est constitué :

- d'un module de synthèse proprement dite qui détermine le genre, le cas, le nombre, le temps des différents termes,
- d'un programme de réarrangement syntaxique,
- et d'un programme de génération morphologique.

3) Le transfert

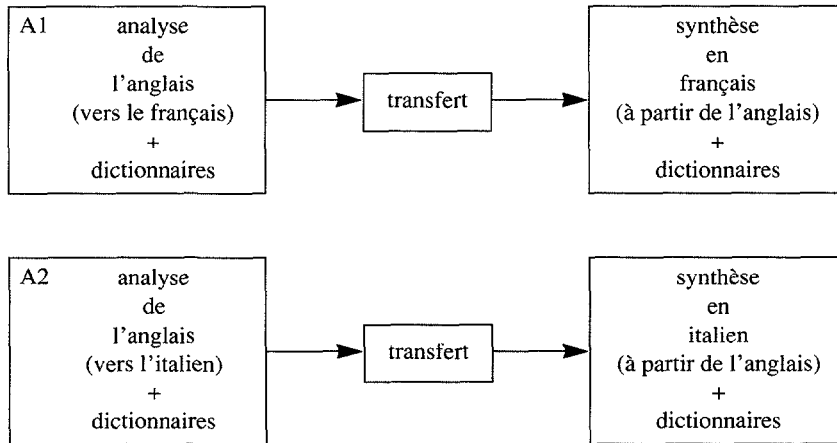
Entre l'analyse et la synthèse, il existe une phase très importante, dite de transfert, s'appuyant particulièrement sur le dictionnaire contextuel dont

nous parlerons ci-après. Composée de routines lexicales permettant le passage de la structure d'une langue à celle d'une autre, elle est donc particulière au couple «langue source — langue cible».

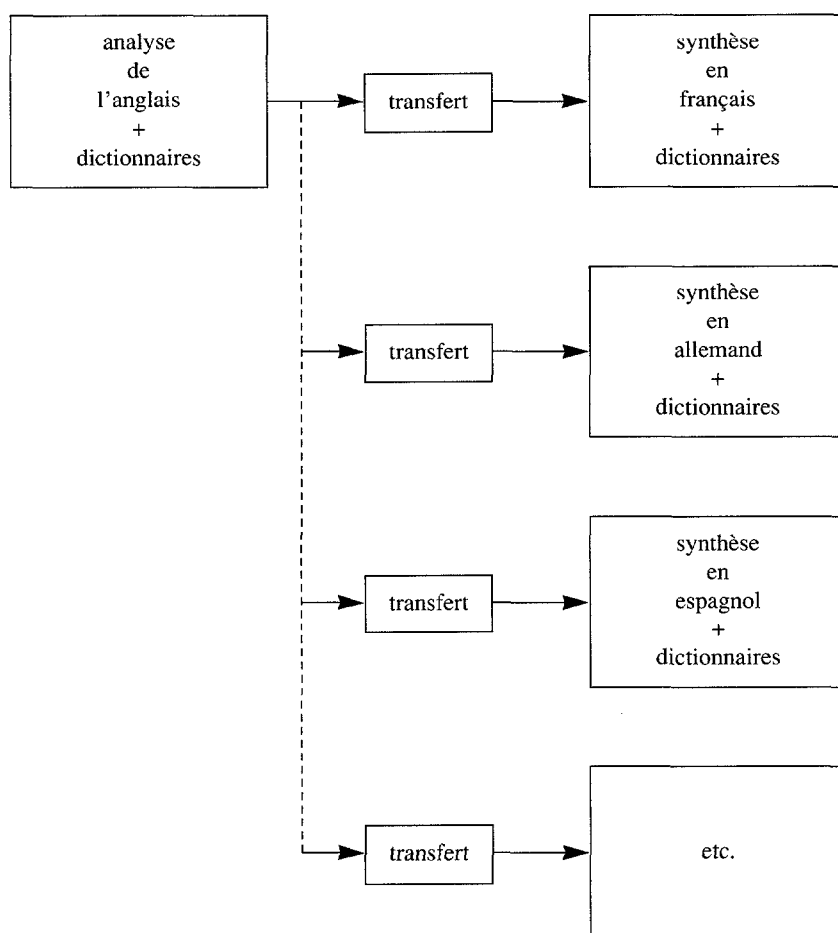
L'un des principaux buts de la modernisation du SYSTRAN a été de rendre le système modulaire et multicible, c'est-à-dire qu'à partir de l'analyse d'une langue source on peut développer plusieurs paires de langues, par exemple, l'anglais vers le français, l'anglais vers l'allemand, l'anglais vers l'espagnol. Il suffit alors d'écrire les 9 analyses de langue source, les 9 synthèses de langue cible ; cependant les 72 programmes de transfert restent toujours nécessaires, car ils représentent les «prises multiples» permettant de décliner une nouvelle paire de langues à partir d'autres déjà existantes.

A ce jour, 23 paires de langues ont été développées à des niveaux de qualité divers. Quinze paires sont opérationnelles et utilisées professionnellement {cf. section 2.2.).

Système monocible



Système multicable



Le SYSTRAN constitue ainsi un ensemble très élaboré et très structuré de programmes et de dictionnaires qui va réaliser une série de va-et-vient sur les mots de chaque phrase, jusqu'à définir l'ensemble des interrelations existant entre tous ces mots.

Il se base, pour ce faire, sur des renseignements concernant les associations possibles entre les mots ou les caractéristiques grammaticales, syntaxiques, sémantiques et prépositionnelles de chaque mot qui lui sont données par les dictionnaires de base.

Après l'analyse, un choix est réalisé entre les multiples sens possibles des expressions grâce à un dictionnaire spécial dit des «expressions à cadre sémantique limité». Celui-ci est, en fait, une base de données de

microprogrammes qui vont décider du sens adéquat de chaque mot en fonction du contexte syntaxique et sémantique.

Une fois la phase de transfert terminée, le programme de synthèse va rédiger à nouveau le texte dans la langue cible en appliquant l'ensemble des règles grammaticales et syntaxiques programmées.

Les étapes parcourues lors de la traduction se présentent comme suit :

- pré-processing,
- expressions idiomatiques invariables,
- dictionnaire principal syntaxique et sémantique,
- dictionnaire des groupes nominaux,
- levée des ambiguïtés homographiques,
- relations entre propositions élémentaires,
- groupes syntagmatiques,
- relations analytiques,
- recherche des énumérations,
- dictionnaires analytiques,
- recherche du sujet,
- relations intrinsèques,
- dictionnaires des expressions verbales et conditionnelles,
- traduction des prépositions,
- structures spéciales,
- régénération grammaticale,
- réarrangement de l'ordre des mots,
- post-processing.

- CONCLUSION :

Nous avons vu tout au long de cette note de synthèse, de multitudes variétés d'architecture de systèmes de traduction automatique, et les domaines de connaissances qui sont utiles et nécessaires à l'élaboration d'un tel système .

Les travaux en linguistique informatique sont en développement continuels malgré l'échec des travaux en traduction automatique avec le début de l'informatique dans les années 1950-60.

Les performances des outils matériels et logiciels des ordinateurs à l'heure actuelle réveillent des travaux en linguistique informatique et en Intelligence Artificielle dans le domaine de la TA.

Une étroite collaboration dans les projets de TA entre linguistes et informaticiens, a permis l'écriture des algorithmes plus performants en matière de connaissances linguistiques et qui rend possible l'automatisation des phénomènes répétitifs dans le processus de la traduction.

Mots, termes, phrases, structures grammaticales ou unités de sens sont autant d'éléments quand ils sont suffisamment répétitifs, que le traducteur humain traite sans même y penser et sans mettre œuvre son génie de la langue.

Les recherches ont envisagé dès lors l'automatisation de cette partie du travail de traduction sans toucher à la partie *sacrée*, réservée à l'homme .

L'exploitation de la répétitivité ne s'arrête pas au niveau lexical ou terminologique, mais s'étend à tous les niveaux du traitement automatique de la langue naturelle.

On retrouve également des régularités dans les différences de style entre les langues.

On peut donc considérer que la TA trouve son fonds de commerce dans la répétition ou plutôt dans une certaine régularité des phénomènes linguistiques: il importe donc bien de connaître les limites d'un système de TA.

Les limites qu'il importe de reconnaître et de respecter sont de deux ordres:

^> *Humain* : le génie linguistique ne s'attaque pas au territoire réservé du génie de la langue propre à l'homme.

^> *Technique* : la description ou le mécanisme de résolution de problèmes en TA peuvent entraîner des inconvénients .

BIBLIOGRAPHIE

ABBOU André ,*Le Japon: Traduction et autoroutes d'information*, La tribune des industries de la langue et de l'information électronique N° 17-18-19, OFIL, 1995. p.76-83.

ABBOU André ,*La théorie du procédé constructif*, La tribune des industries de la langue et de l'information électronique N° 17-18-19, OFIL, 1995. p.35-36.

ABBOU André, *Traduction Assistée par Ordinateur - Actes du séminaire international (Paris, Mars 1988)*, Ed. DAICADIF, 1989. 234 p.

AL-ABED Lamia, *Traduction Assistée par Ordinateur de l'anglais vers l'arabe*, Rapport de synthèse sur le laboratoire TAO de 1990-1992 - Département Arabisation & Communication Homme-Machine, IRSIT Tunisie Déc.1992, p.10 .

AL-ABED Lamia, *Elaboration d'un système de traduction de phrases de l'anglais vers l'arabe*, Mémoire d'ingénieur concepteur en informatique, FST, 1990, p.205 .

ABEILLE Anne, *les nouvelles syntaxes*, Armond Collin, 1993, 256 p.

ABU AL-CHAY Najim, *Un système expert pour l'analyse et la synthèse des verbes arabes dans un cadre d'enseignement assistée par ordinateur*, Thèse de doctorat, Université Claude Bernard , Lyon 1, 1988. 139 p.

ALLEN James , *Natural language understanding* ,2nd Ed. , The Benjamin / Cummings publishing Company INC, 1995 . 654 p.

AMANO S. (& HIRAKAWA H. , NOGAMI H. , KUMANO A.), *The TOSHIBA machine translation*, Future Computing System vol.2 N°3, Oxford University Press and Maruzen Company Limited, 1989. p. 227-246.

ARCHIBALD Michiels, *Traitement du langage naturel et Prolog*, HERMES Ed. , 1991.

ARNOLD Dong (& DES TOMBE Louis) , *Basic theory and methodology in EUROTRA*, Machine translation theoretical and methodological issues, University Press of Combridge, 1987. p. 114-134.

BALLON S . (& SARRON C.), *Industries de la langue* - Centre Français du commerce extérieur - Direction des industries et services, Ed. du CFCE, Mai 1994. 76 p.

BIERIN E. (& MOULIN A. , PICHAULT F.), *Les industrie de la langue: un marché en devenir*, OWIL,1990. 98 p.

BILANGE Eric, *Dialogue personne-machine - Modélisation et réalisation informatique*, Edition HERMES, 1992. 192 p.

BOLC Léonard, *Natural language parsing systems*, Springer-Verlag: *Symbolic Computation A.I, Germany*, 1987. p. 60-110

BOUALEM Abdel-Malek, *ML-TASC Système de conversion de formalismes de langages techniques et scientifiques dans un environnement à syntaxe contrôlée et à raisonnement limité-Traduction Automatique Multilingue*, Thèse de docteur en science, Université Sophia Antipolis, 25 Nov. 1993. 262 p.

BOULICAUT J-F , *Vers une programmation grammaticale à large spectre - Application à la reconstruction de programmes avec le métacompileur STARLET/GL*, Thèse de doctorat en science, INSA Lyon, 1992. 319 p.

BOURBEAU Laurent (& ISABELLE Pierre), *TAUM-AVIATION: Its Technical Features and some experimental results*, Computational Linguistics Vol: 11 N° 1, 01-03 1985. p. 18-27.

CARBONELL Jaime G. (& TOMITA Masaru), *Knowledge-based MT the CMU approach, Machine translation theoretical and methodological issues*, University Press of Combridge, 1987. p. 68-89.

CHURCH K. W. (& HOVY E. H.), *Good applications for CRUMMY machine translation*, *Machine Translation 8*, Kluwer Academic Publishers Netherlands, 1993. p. 239-258.

HUTCHINS, W. J.: *Machine Translation. Past, Présent, Future*, New York, Ellis Horwood, 1986. 382 p.

CULLINFORD Richard E.(& ONYSHKEVYCH Boyan A.), An experiment in lexicon-driven MT , *Machine translation theoretical and methodological issues*: University Press of Cambridge, 1987. p. 278-301.

Chapitre IV : Quelques difficultés liées à la T.A

On évoquera dans ce chapitre quelques difficultés majeures qui entravent l'aboutissement à une traduction automatique fiable dont la délimitation des unités de traduction et la conception de dictionnaires.

I.Délimitation des unites de traduction :

Introduction :

La traduction, par essence, opère sur l'idée, le sens. Il lui faut, donc, un élément de base qui lui permet de rendre l'idée, de décomposer la phrase en segments porteurs de sens et entièrement restituables. C'est l'unité de traduction. Selon Newmark 'l'unité de traduction s'applique normalement à des unités de la langue de départ qui peuvent se recréer dans la langue d'arrivée sans addition d'autres éléments de sens de la langue de départ' et, pour lui, l'unité de traduction est ' toujours le plus petit segment de l'original qui fournit un équivalent acceptable pour un segment du texte traduit'. La notion d'acceptabilité étant trop subjective, il serait préférable d'opter pour la définition proposée par Vinay et Darbelnet : ' l'unité de traduction est le plus petit assemblage de mots qui contribuent à l'expression d'un seul segment de pensée et dont le degré d'interdépendance est tel qu'ils ne peuvent se traduire séparément sous peine de modifier ou d'annuler le sens du message'.

S'il y a une étape importante dans le processus de traduction automatique est celle relative à la délimitation de cette unité de traduction, c'est-à-dire le fragment de la phrase qui reste irréductible sémantiquement renvoyant à un sens unique et un seul. L'unité de traduction serait l'entité suffisamment identifiable et reconnaissable par la machine pour être traitée, représentée et transférée. Si cela apparaît pour l'humain une évidence, ce n'est pas le cas pour une machine

à qui tout n'est qu'une suite de caractères. Le problème est, donc, de les identifier, de les isoler et de déterminer le dénominateur commun qui leur soit applicable pour les faire passer d'une langue à l'autre.

La recherche de telles unités dans une phrase reste compliquée car il s'agit d'un concept reconnaissable de la signification qui n'est pas toujours donné par le mot et peut être un ensemble ou paquet de composants de signification. Ainsi Larson citant Barnwell ' " dans une langue donnée, l'unité de concept habituellement, mais nullement toujours, est représenté par un mot; il peut également être représenté par un morphème, ou par une expression idiomatique, ou par tonalité, ou par ordre des mots. Des concepts sont identifiés dans une langue donnée selon le principe du contraste et de la comparaison dans le système de cette langue ".

Dans ce sens, l'unité de traduction ne peut être réduite au seul mot et peut en comprendre plusieurs. Sitôt l'unité identifiée, il serait facile de chercher un équivalent dans la langue d'arrivée si, bien sur, ce dernier existe. Quel est donc le moyen pour délimiter l'unité de traduction ou unité sémantique ?

I.1.le mot est-il une unité valable ?

Parfois, le mot peut-être une unité valable et suffisante, surtout dans les trois systèmes linguistiques qui nous intéressent (Arabe, Français et Anglais).

Ainsi dans la phrase simple :

Il mange une pomme.

peut être délimitée en trois unités distinctes : / Il / mange / une pomme /. Chacune d'elle peut exister indépendamment des deux autres et dont le sens unique est suggéré par elle et elle seule. Dans ce cas, une traduction vers l'arabe se fera en remplaçant chaque unité par son équivalent immédiat, c'est-à-dire :

/Il/ → هو

/mange/ → يأكل

/une pomme/ → تفاحة

Ensuite un simple réaménagement morphosyntaxique dans la langue d'arrivée aboutirait à générer une phrase cohérente, intelligible et correcte. Ceci reste valable, si l'unité, le mot, n'est pas polysémique ou ambigu. Si le mot est polysémique ou ambigu, on utilisera, une méthode de désambiguïsation.

De même, dans la phrase anglaise suivante :

He usually takes his car to go to work.

La délimitation en unités de traduction serait:

/ He / usually / takes / his / car / to / go / to / work

dont chacune est monosémique et non ambiguë. La règle de correspondance, par le biais d'un dictionnaire approprié suivi d'un réaménagement morphosyntaxique en langue arabe générerait une phrase traduite en langue cible, en l'occurrence la langue arabe. Sauf que dans cet exemple, le verbe 'to take', bien qu'il ne soit pas ambigu, nécessite une 'réinterprétation' en langue arabe dans le sens où une traduction littérale du verbe en arabe aboutirait à un usage non conforme. 'to take' doit être interprété dans le sens 'utiliser' et non 'prendre'. Si pour le traducteur humain, cet écueil peut être facilement éludé, pour la machine, cela reste insurmontable vu les limites du dictionnaire électronique interne de correspondance entre la langue de départ et la langue d'arrivée qui ne peut contenir toutes les subtilités possibles de chaque langue ainsi que le génie suggestif de chaque mot en dehors de sa polysémie ou ambiguïté. Ici, bien que l'unité de traduction, le mot, est bien identifié, la machine ne peut s'affranchir de la difficulté de lui trouver un correspondant exact dans la langue d'arrivée, comme le constate un linguiste 'dans la phrase, les mots se déterminent les uns les autres ; une sélection s'opère ainsi entre leurs diverses significations possibles... C'est dire qu'en fin de compte un mot n'a pas de sens en soi : il n'a de sens que dans et par un contexte'.

Il s'avère, alors, que le mot peut être appréhendé par la machine comme unité de traduction dans les deux langues (anglaise et française) mais sous certaines conditions.

I.1.a. le mot ne fait pas partie d'un mot composé :

Examinons les exemples suivants :

He asks the **shop assistant**. (1)

The **shop** is closed. (2)

His **assistant** advises him to not intervene. (3)

Les deux mots ‘shop’ et ‘assistant’ dans (2) et (3) sont, en effet, des unités de traduction, par contre, dans (1), aucun des deux ne l’est. Plus encore, les deux forment une autre unité de traduction (un mot composé). Pour que la machine puisse décider si ‘shop’ est, par exemple, une unité de traduction ou pas et opter pour la traduction adéquate, elle analyse le mot suivant, sachant que le mot composé ‘shop assistant’ est contenu dans le dictionnaire électronique interne.

Ceci passe, donc, par répertorier tous les mots composés de la langue française et anglaise de telle façon que la machine puisse les consulter pour décider si le mot est ou pas une unité traduction ou fait partie d’un mot composé.

Sauf que cette solution n’est pas toujours d’un bon secours parce qu’il arrive qu’un mot composé ne soit pas connu de la langue, pas assimilé comme telle et donc pas répertorié ; son statut de ‘mot composé’ n’est perceptible que dans son usage dans la phrase :

The money must be used to **good purpose**.

Ici ‘good purpose’ est une seule unité de traduction, genre ‘mot composé’ bien que grammaticalement il joue le rôle d’un adverbe.

Un autre exemple :

As a **repeat offender** he faces a maximum penalty of three years’
imprisonment.

où ‘repeat offender’ est une unité de traduction qui doit être prise pour telle par la machine et non pas fragmentée en deux ‘unités mots’, le nom ou verbe ‘repeat’ et le nom ‘offender’.

Un autre type de mots composés spécifique à l’anglais :

‘I’m afraid this will do more harm than good.’ I said. She **noded agreement**.

L’unité de traduction ici joue le rôle d’un verbe dans le sens ‘acquiesça’.

On souligne le même fait ou phénomène en ce qui concerne la langue française bien qu'en anglais cela est plus fourni et omniprésent. En effet, la fréquence des mots composés dans la langue anglaise est plus prépondérante que dans la langue française surtout dans les textes scientifiques et techniques.

Longtemps je me suis couché de **bonne heure**. (1)

Il a eu une **bonne** note. (2)

Il est arrivé à l'**heure**. (3)

'bonne heure' dans (1) est une seule unité de traduction ou une seule unité sémantique car les deux mots s'associent pour donner un seul sens 'tôt' à l'inverse des deux exemples qui suivent où 'bonne' et 'heure' sont une unité de traduction simple.

On peut citer plein d'autres exemples. Mais ce qui peut être souligné est le fait qu'un mot n'est pas toujours, en anglais et en français, une unité de traduction. Ensuite, les mots composés ne sont pas toujours ceux qui existent dans les dictionnaires classiques, d'où la nécessité de créer au préalable un dictionnaire électronique interne regroupant tous les 'mots composés' auquel la machine se réfère pour distinguer l'unité de traduction adéquate, un mot ou une association de plusieurs.

I.1.b. le mot ne fait pas partie d'une expression idiomatique :

Les expressions ou constructions propres à une langue et dont le sens ne peut se déduire de l'analyse de sa structure et du sens de chacun de ses éléments restent intraduisibles littéralement d'une langue à l'autre. C'est pourquoi le mot dans de pareilles expressions doit être discerné pour voir s'il est une unité de traduction par lui-même ou fait partie d'une expression idiomatique.

Considérons les exemples suivants :

She **looks after** her baby. (1)

She **looks** gloomy. (2)

Dans (2) 'looks' est une unité de traduction représentée par un seul mot, par contre dans (1) 'looks' est associé au mot suivant 'after' qui génère une seule

unité de traduction bien différente dans les sens. Pour éviter cette confusion dans la délimitation , il convient de stigmatiser ces idiomes dans des répertoires spéciaux. Cela paraît plus simple car l'ensemble des idiomes anglais est connu et est exhaustif.

Généralement, en langue anglaise, les formes idiomatiques sont associées aux verbes suivis d'un ou plusieurs prépositions :

She **looks after** her baby. (1)

We must **look ahead** to the future now. (2)

He **looks around** at his team. (3)

He **looks around** for his friend. (4)

They spent the morning **looking around** London. (5)

Look at this coat book! (6)

He **looks back** at his friend. (7)

Let's **look back** to the year 1964. (8)

From the hilltop she **looked down** on the city. (9)

He **looks for** his shoes in the bathroom. (10)

she **looks forward** to going on holiday. (11)

I'll **look in** again tomorrow. (12)

look in on the baby and check she's still asleep. (13)

They **look into** the water issue. (14)

I was forced to **look on** as the house was ransacked. (15)

The room **looks onto** the garden. (16)

you must **look out** for snakes. (17)

The window **looks out over** the sea. (18)

He **looks over** the contract. (19)

She **looked round** to see who it was. (20)

He **looks through** the book. (21)

They **look to** him for leadership. (22)

He **looks up** to the clouds. (23)

They **look up to** their leader. (24)

De prime abord, on constate que seul pour les idiomes relatifs au verbe ‘to look’, on en retrouve 24 ! Ce qui reste impressionnant quand on sait que presque tous sont polysémiques. Autre remarque, les idiomes ne sont pas toujours constitués d’un verbe suivi d’une seule préposition (les exemples 18 et 24). Ce qui complique la tâche de la machine. Où doit-elle arrêter le découpage en unité de traduction dans les expressions idiomatiques ? Dans les exemples précédents, si dans (17) et (18), cela reste possible. En effet ‘look out for’ n’est pas répertorié comme expression idiomatique, par contre ‘look out over’ en est un. Ce qui n’est pas le cas dans les exemples (23) et (24) où l’ambiguïté dans le découpage est criarde. Serait-ce ‘to look up’ ou ‘look up to’ car tous les deux sont répertoriée comme étant des expressions idiomatiques. De surcroît, dans l’exemple (24), si on opte pour l’un ou l’autre des découpages, cela donne deux sens différents, tous deux exacts et possibles. Lequel opter ? Dans l’exemple (23), ce problème peut être facilement évité en appliquant un algorithme approprié (une personne ne peut admirer les nuages) en faisant la restriction que ‘to look up to’ ne peut avoir pour objet un ‘non humain’.

Autre obstacle dans la délimitation de l’unité de traduction, en ce qui concerne les expressions idiomatiques en anglais, est que leurs constituants ne sont pas toujours contigus. Examinons les exemples suivants :

He **looks** his number **up** in the phone book.

The expert must **look** the car **over** before you buy it.

Ce phénomène propre à l’anglais altère fortement les performances de la machine car il la rend inopérante et nécessite des algorithmes très complexes. Il faut, en effet, analyser toute la phrase pour vérifier s’il existe une préposition qui peut être liée au verbe, ce qui n’est pas toujours le cas. Et là la machine doit décider au risque de donner une traduction erronée.

Cette ambiguïté qui entoure la délimitation des unités de traduction reste palpable dans les deux exemples suivants :

To look to somebody. (1)

To look to him for leadership. (2)

‘to’ dans (1) est une préposition est donc une unité de traduction à part. Dans l’exemple (2), ‘to’ fait partie intégrante de l’unité de traduction (l’expression idiomatique) ‘to look to’. Lever cette ambiguïté serait impossible par un algorithme car l’environnement grammatical, syntaxique et de connaissance est identique.

I.1.c. le mot ne fait pas partie d’une conjonction ou locution adverbiale.... :

Considérons les exemples suivants :

He speaks a **good** English. (1)

The match is **as good as** lost. (2)

She left him **for good**. (3)

Il apparaît bien que dans (1) ‘good’ est une unité de traduction qui peut être traitée par la machine indépendamment des autres unités. Ce qui n’est pas le cas dans les deux exemples qui suivent. Dans le deuxième, il ne peut être dissocié de ‘as’ avant et après, dans le troisième de ‘for’ avant. Dans (2), le sens serait ‘quasiment’ et dans (3) ‘pour toujours’. Donc ‘as good as’ et ‘for good’ sont deux unités de traduction qui doivent exister en tant que telles et introduites, au préalable, dans un dictionnaire électronique interne de correspondance.

Et là l’ordinateur serait confronté à une double difficulté ; pouvoir distinguer la conjonction de la locution adverbiale, et l’ambiguïté de la locution adverbiale elle-même :

He is **as good as** his brother.

L’ambiguïté est, ici, générée par le découpage en unités de traduction au sein de la phrase pour retrouver l’équivalent exact dans la langue d’arrivée et qui pose un problème lors de la réalisation d’un dictionnaire interne des équivalences.

Remarquons que si ‘good’ est au pluriel (adjonction d’un ‘s’ à la fin du mot ce qui est dans ce cas un mot pluriel singulier), la confusion est levée.

On est pas encore au bout de nos peines concernant l'usage de 'as' et sa perception par la machine :

He left early **so as** to be on time.

Ici 'so as' est une seule unité de traduction, par contre dans:

She worries **so as** always.

'so' et 'as' sont deux unités de traductions distinctes bien qu'elles soient contiguës. Il serait difficile, encore une fois, à la machine de faire le découpage adéquat. Des algorithmes doivent être prévus pour éviter cela. La complexités des phrases où apparaissent des cas pareils et leurs multitudes rend la tâche difficile dans le sens où on ne peut pas adopter à chaque problème un algorithme qui lui est propre sachant que chaque phrase comporte ses propres caractéristiques syntaxiques, stylistiques, contextuels et grammaticales. En un mot pour faire un algorithme efficace, il faut que la phrase soit reproductible à une certaine échelle de par son environnement et ses spécificités. C'est-à-dire que le modèle et sa représentation peuvent générer d'autres phrases semblables dans leurs caractéristiques.

Donc, comme on le constate la difficulté du découpage en unité de traduction se pose avec acuité.

Aussi, comme dans le cas des expressions idiomatiques, les locutions adverbiales ne sont pas toujours forcément contiguës.

He is **so** fat **that** he can't get in. (il est tellement gros qu'il ne peut pas entrer)

It is **so** worded **that** helps him to understand. (c'est rédigé de telle façon qu'il puisse comprendre)

She wrote the instructions **so that** they'd be easily understood..

Dans les deux derniers exemples, il s'agit de la même unité de traduction bien que 'so... that' et 'so that' n'apparaissent pas sous la même forme. Le premier exemple démontre que 'so...that' peut être deux unités de traductions à l'inverse du deuxième exemple.

I-1-d- le mot ne fait pas partie d'une tournure, d'un proverbe ou d'une expression consacrée :

Il est évident qu'un mot, s'il fait partie d'une tournure, d'un proverbe ou d'une expression consacrée, n'est plus une unité de traduction en soi. Son sens se trouve englouti et c'est la tournure, le proverbe ou l'expression consacrée qui deviennent significatifs.

I don't **eat** meat.

I'll soon have him **eating** out of my hand. (Bientôt j'en ferai ce que je voudrai)

He **eats** his heart out. (il souffre en silence).

On constate que le mot 'to eat' n'est plus une unité de traduction dans les deux derniers exemples et il n'apparaît même pas, ni implicitement ni explicitement. La machine ne peut faire un découpage en unités de traduction pour retrouver les entités de signification. Ici c'est l'ensemble qui a une signification. Dans ce cas, il est indispensable d'enregistrer tous les proverbes, les tournures et expressions consacrées.

He **sits** on the floor.

He **sits** for the exam.

Si dans le premier exemple 'sit' est une unité de traduction à part, dans le deuxième, la machine ne peut prendre celui-ci ('sit') comme telle. Alors comment peut-elle déterminer si le mot est ou pas une unité de traduction ? La seule solution serait d'enregistrer toutes les expressions consacrées avec leurs traductions complètes. L'ordinateur pourra directement y accéder et restituer la traduction.

Dans ce sens Van Hoof (5), lui qui nie l'existence du mot comme unité de traduction, dresse un tableau exhaustif de l'agencement des mots en faisant une classification d'ordre (a) numérique (basée sur le nombre des mots), (b) fonctionnel (basée sur la fonction grammaticale), (c) contextuel (basée sur le rôle des unités dans le contexte) et (d) cohésionnel (basée sur le degré de cohésion des mots à l'intérieur des unités). Et ce sont les deux dernières classifications qui sont intéressantes et qui siéent le traitement automatique des unités de traduction et qui méritent qu'on se penche dessus.

I.2. Classification contextuelle :

Selon que, dans le contexte, l'unité circonscrit un segment de pensée, en exprime le ton, ou marque l'enchaînement de divers segments entre eux. Il distingue trois catégories :

- les unités sémantiques, dont les termes participent à un même sens :
e.g. out of hand, to rise to one's feet, fourth-class mail....
Au pied de la lettre, prendre la mouche, à Cœur ouvert....
- les unités tonales, dont les termes participent à une même intonation :
e.g. not half !, don't tell me !, bet you I will!
tu parles!, mince (alors) !, t'occupes pas !....
- les unités articulatoires, dont les termes participent à une même articulation
e.g. with respect to, in spite of...
par contre, ainsi donc...

I.3. Classification cohésionnelle :

Le degré de cohésion qui lie les termes à l'intérieur des unités de traduction permet de classer celles-ci en deux catégories distinctes, les groupes unifiés et les groupes par affinité :

(a) les groupes unifiés sont des unités dont les termes forment bloc au point d'être inséparables sous peine de contresens en cas de construction mot à mot.

On trouvera dans cette catégorie :

- des locutions figées :
e.g ; all of a sudden, on all fours....
En un tour de main, se rendre compte, de pied en cap...
- des expressions idiomatiques :
e.g. ; to talk to the wind, to be in the doldrums, to have a sweet tooth....
- des dictons et proverbes:

e.g. dog does not eat dog, to make hay while the sun shines, birds of a feather flock together...

à bon chat bon rat, chat échaudé craint l'eau froide...

- des tours qui relèvent des niveaux de langage :

e.g. full of beans (familier), to spill the beans (populaire), to hit the hay (argotique), direct light reflex (spécialisé)...

perdre le nord (familier), passer à tabac (populaire), travailler du chou (argotique), coefficient d'épuration (spécialisé)...

(b) des groupements par affinité sont des unités dont les termes s'associent plus librement. Leur cohésion, même occasionnelle, reste cependant perceptible pour constituer une unité sémantique et donc, une unité de traduction.

On y trouve une très grande variété de locutions, dont on peut citer, à titre d'exemple :

- les locutions d'intensité : dans lesquelles le nom appelle un adjectif ou une locution adjectivale qui l'intensifie. Elles représentent une seule unité de traduction dans le sens où on ne peut pas faire une traduction littérale (l'adjectif et le nom vont ensemble). Elles sont assimilées à des 'pseudo' unités de traduction.

E.g. square refusal (refus catégorique), out-and-out liar (fieffé menteur), white rage (colère bleue)...

Démenti formel, misère noire, perte sèche, célibataire endurci....

Dans ce sens, Van Hoof stipule, qu'en anglais, l'intensification du nom peut se faire par un tour adjectival construit sur un autre nom, du type a(n) of a (n), construction que l'on retrouve le plus souvent en français sous la forme un(n) de (n) :

e.g. a hell of a din, apeach of a girl, a devil of a wind...

un vacarme d'enfer, une fille de tonnerre, un vent du diable...

justifiant cela par le fait que c'est toujours le second nom qui est prépondérant et le mot qui suit serait un adjectif. Il cite l'exemple :

‘a love of a child’ qui serait, si traduit selon sa classification, ‘ un enfant de l’amour’. Ce qui serait un contresens.

- des locutions telles que : muet comme la tombe, léger comme une plume où c’est l’adjectif qui reçoit l’intensification. Mais, très souvent, le groupement par affinité qui leur correspond dans l’autre langue ne coïncide pas sur la marque d’intensification :

e.g. solemn as a judge (sérieux comme un pape), dumb as an oyster (muet comme une carpe), cross as two sticks (d’une humeur de chien), red as a boiled lobster (rouge comme une pivoine), drunk as a Dutchman (saoul comme un polonais)...

Ennuyeux comme la pluie (as dull as ditchwater), bête à manger du foin (as stupid as an owl), fort comme un turc (as strong as a horse), bête comme chou (as easy as shelling peas), jaune comme un citron (as yellow as a guinea)....

En anglais, l’intensification de l’adjectif peut s’obtenir plus simplement par anté position d’un nom ou d’un adjectif, verbal ou non :

e.g. dirt cheap (vachement bon marché), stone deaf (sourd comme un pot), red hot (chauffé à blanc), soaking wet (trempé comme une soupe), raving mad (fou à lier)...

- dans les locutions d’intensité construites sur un participe passé, les marques d’intensification sont également très souvent différentes d’une langue à l’autre :

e.g. stark naked (nu comme un ver), caught red-handed (pris la main dans le sac), bored stiff (ennuyé à mourir), chilled to the bones (transi jusqu’à la moelle), dressed up to the nines (tiré à quatre épingles)...

tué net (killed on the spot), raide mort (stone dead), battu à plate couture (beaten out and out), trempé jusqu’aux os (wet to the skin), usé jusqu’à la corde (threadbare)...

- il en va de même pour les locutions construites sur un verbe, où la marque d’intensification peut aller jusqu’à englober la totalité de l’unité (to bring down the house, par exemple):

e.g. to worry oneself to death (se tourmenter à s'en rendre malade), to blush to the roots of one's hair (rougir jusqu'au blanc des yeux), to beat someone black and blue (battre quelqu'un comme platre), to rain cats and dogs (pleuvoir à seaux), to win hands down (gagner haut la main)....

Applauder à tout casser (to bring down the house), dormir à poings fermés (to sleep like a log), manger comme quatre (to eat like a wolf), rire aux éclats (to laugh till one's sides ache), crier à tue-tête (to shout at the top of one's voice)....

- des locutions adjectivales et adverbials:

e.g. hors de prix (prohibitive), bon marché (cheap), hors pair (unrivalled, peerless), bon teint (fast ; staunch), hors de propos (irrelevant), à tort (wrongly), de bon gré (willingly), mine de rien (unobtrusively), de fond en comble (utterly), mal à propos (unseasonably)...

up to date (moderne ; actuel), out of the way (écarté ; perdu), out of focus (brouillé), touch and go (risqué, hasardeux), dyed in the wool (endurci; convaincu), right away (immédiatement), straight off (d'emblée), in the meantime (entretemps), now and again (parfois), by the way (incidemment)...

- des locutions de qualification:

de nombreuses locutions ne sont constituées que d'un nom et d'un adjectif mais la marque de qualification apportée par ce dernier s'est effacée devant l'unité nouvelle créée :

e.g. walking lady (figurante), white collar (employé), brass hat (officier supérieur), red tape (bureaucratie), hot dog (petit pain fourré d'une saucisse chaude)....

Petite main (handy girl), grand magasin (department store), beau parleur (glib tongue), tête brûlée (dare-devil), faux frère (blackleg, double-crosser)...

Ce qu'on entend, aussi, par expressions consacrées, les néologismes et appellations nouvelles devenues des faits de langue établis. Ceci concerne, surtout, le registre scientifique qui ne cesse de se développer jour après jour.

II- Conception des dictionnaires électroniques :

Nous nous intéresserons, ici, à la consultation des dictionnaires électroniques lors du processus de traduction automatique. L'exercice de traduction nous a tous placés dans une situation embarrassante : ayant recherché dans un dictionnaire un mot inconnu, celui-ci propose plusieurs solutions. Laquelle choisir ? c'est bien sur la compréhension des autres éléments du texte qui permet le choix, encore faut-il que ces autres mots aient déjà reçu une traduction correcte. C'est là qu'un jeu de devinettes intervient, du aux différentes combinaisons possibles et probables, au cours duquel une grande ingéniosité ne peut éviter faux-sens et contre-sens. Qu'en est-il de la machine qui par nature ne dispose pas des mécanismes que met en jeu le traducteur humain : sa souplesse de raisonnement par tâtonnements multiples et son évaluation de la plausibilité des interprétations ?

Un ordinateur peut explorer les contextes à droite et à gauche d'un mot en vue de résoudre les ambiguïtés, mais pour ce faire, il doit disposer d'informations strictement codifiées qui auront été introduites au préalable dans son dictionnaire et dans sa grammaire. Ces deux composants sont cruciaux dans le traitement automatique des textes.

Toutefois, le courant majoritaire des recherches en traitement automatique des langues y compris en traduction automatique assistée (TAO) ou non, ne prend jamais en compte le composant du dictionnaire. Ce n'est que dans les produits commerciaux qui vont de SYSTRAN à EUROLANG que des dictionnaires ont été intégrés aux systèmes de TAO. Il en va de même pour les grammaires. C'est donc un paradoxe de constater que pendant quarante ans, les recherches sur l'analyse automatique du texte et la traduction se sont poursuivies sans dictionnaires ni grammaires et sans aucun programme scientifique de construction de ces composants. Les approches de recherches ont pris une orientation différente, elles font intervenir d'autres activités et procédures parmi lesquelles :

II.1.L'étiquetage :

Les travaux sur l'étiquetage des textes ont pour but d'attacher aux mots d'un texte leurs attributs grammaticaux. Les procédures mettent en jeu des analyses morphologiques et des textes sur les contextes : par exemple la détection en français du suffixe 'ations' permet d'affirmer qu'il appartient à un substantif féminin pluriel. L'exemple de 'cations' substantif masculin pluriel est une exception, celui de 'rations' est plus complexe et plus général : le mot est ambigu, est-ce le verbe 'rater' ou bien le pluriel du nom 'ration'. Le verbe 'rations' est à la première personne du pluriel, on alors la quasi-certitude que le pronom sujet 'nous' apparaîtra dans son voisinage immédiat. Un test formel simple lèvera donc cette ambiguïté.

La finalité de ces travaux n'est pas claire, car même si les résultats de l'étiquetage étaient de bonne qualité, l'information attachée aux mots d'un texte sera limitée aux seules valeurs grammaticales. Or de nombreuses autres informations sont indispensables à l'analyse utile des textes.

Il a bien été proposé que des textes étiquetés pourraient faire l'objet de statistiques sur des séquences de catégories grammaticales, ce qui permettrait éventuellement de dégager des régularités qui faciliteraient l'analyse automatique. L'étiquetage a récemment fait l'objet d'une généralisation spectaculaire : en étiquetant des textes bilingues, les statistiques pourraient dégager des régularités de traduction et donc intervenir en TA ou TAO.

Les résultats obtenus aujourd'hui par l'étiquetage d'une seule langue sont bien maigres, on ne peut guère s'attendre à des améliorations dans le cadre de deux langues couplées par une traduction.

II.2. La modélisation :

Les recherches sur les modèles linguistiques adéquats au traitement automatique des langues sont nombreuses et anciennes. Elles n'ont jamais mis en jeu de dictionnaires, ni de grammaires. Les modèles ont fait l'objet d'une élaboration logique raffinée et il en existe un grand nombre, chacun avec un grand nombre de variantes, aucun ne met en jeu plus de quelques

dizaines ou centaines d'éléments lexicaux et plus de quelques dizaines de règles les combinant. Tout l'effort porte donc sur des définitions de formalisme sans préoccupation de leur adéquation empirique. On suppose qu'un modèle élaboré pourra être chargé d'un bon dictionnaire du commerce et qu'il fonctionnera en toute généralité. Bien sur, le dictionnaire de commerce n'est pas adapté à la machine, il devra donc subir un formatage. Un pont sera établi entre les deux types de dictionnaires, grâce à des travaux qui, s'ils aboutissent, permettront une récupération automatique d'informations.

En apparence, ces deux approches présentent une certaine logique, mais on va voir qu'un simple examen des besoins en dictionnaires pour ordinateurs montre l'absence de pertinence de ces tentatives pour les problèmes et explique ainsi l'absence totale de résultats.

II.3.Observations sur les besoins en dictionnaires :

- Mots simples :

Des expériences faites sur des corpus variés donnent une première indication sur la couverture des dictionnaires du commerce dont les entrées sont des mots simples (Hachette, Larousse, petit Robert). Ces dictionnaires comportent de l'ordre de 80000 entrées simples. L'équivalent électronique contient un peu plus de 90000 entrées. Chacune de ces entrées comporte des codes de flexion (conjugaison, mise au féminin, au pluriel) qui engendrent plus de 700 000 formes fléchies, ce sont les formes rencontrées dans les textes. La procédure de flexion attache à chaque forme ses valeurs grammaticales. L'expérience qui consiste à consulter ce dictionnaire pour un corpus de textes techniques, littéraires et journalistiques, indique des taux de couverture variant entre 70% et 80%. Autrement dit, si on définit un mot formel comme une séquence de caractères d'un texte limitée par deux séparateurs (blancs et ponctuation), une proportion d'un mot du

texte sur cinq ne figure pas dans le dictionnaire. Ces échecs de consultations sont de différents types :

- Mots numériques (chiffres arabes ou romains),
- Noms propres,
- Mots absents des dictionnaires : par exemple, les mots ‘inrenouvelable’ et ‘inrenouvelabilité’ ne figurent dans aucun dictionnaire du commerce, leur recherche suppose une analyse morphologique intuitive du locuteur qui reconnaitra sans trop de peine la base verbale ‘renouveler’.

La recherche des mots d’un texte dans un dictionnaire purement grammatical apporte une information qui sera utilisée pour effectuer une analyse plus approfondie du texte. Mais compte tenu de la proportion des mots non reconnus (entre 20% et 30%), les chances de réussite pour les analyses de phrases sont considérablement réduites.

Les remèdes existent mais ne sont pas simples à mettre en œuvre :

- Les nombres peuvent être aisément reconnus individuellement par leur code ASCII. Pour les chiffres romains, la procédure consistera à en écrire une grammaire, ce qui fournit une bonne approximation du problème. Mais ces procédures simples ne permettent pas de déterminer le rôle des numéraux dans la phrase ;
- Les noms propres pourraient faire l’objet de dictionnaire ;
- Les formations par préfixation et suffixation peuvent être analysées et une solution pratiquement complète du problème a été donnée par D.Clémenceau en 1994.

L’adjonction de ces procédures devrait conduire à un taux de couverture des textes proche de 100%. Mais il manque à ce stade de nombreuses informations lexicales indispensables à l’analyse des textes et à leur traduction.

- Mots composés :

Il est bien connu que les formes idiomatiques ne peuvent être traduites mot à mot (sauf exception). Ces formes sont de toute nature :

- Phrases entières : la moutarde lui monte au nez,
- Adverbes : de temps à autre, par les temps qui courent,
- Adjectifs : fidèle au poste, sur et certain,
- Noms : pomme de terre, microscope à effet tunnel.

Les catégories phrases, adverbe et adjectif ne correspondent pas à des effectifs nombreux. Ainsi, en langue française :

- On recense 50 000 phrases figées, nombre à comparer avec les 12 000 phrases libres du lexique des verbes français (A.Guillet, C.Leclère, 1992) ;
- On évalue 7 000 adverbes figés (M.Gross, 1990).

Ce sont les noms composés qui constituent l'essentiel du composant lexical des langues : les termes techniques font l'objet d'une création continue et l'on évalue leur nombre à plusieurs millions d'entrées.

A ces problèmes d'effectifs, s'ajoutent des problèmes de précision informatique. Ainsi, l'emploi ou pas des majuscules et celui des traits d'union engendre des graphies différentes pour le même mot composé. Aussi l'usage d'un graphe ou d'une représentation s'avère nécessaire. Ce graphe (ou cette représentation) a des avantages importants :

- Il évite des répétitions : le graphe est plus compact que la liste des graphies possible, il est plus lisible dans le sens où il met en évidence les causes de variations ;
- Il explicite la relation d'équivalence qui lie les différentes formes, il permet donc de choisir un représentant unique pour la classe ;
- Il se traduit mécaniquement en un algorithme de reconnaissance de formes dans les textes.

La même représentation peut-être utilisée pour représenter d'autres types de variantes d'un terme donné (variantes morphosyntaxiques et sémantiques). Ainsi par exemple l'expression 'insuffisance de la glande thyroïde' peut avoir plusieurs (4) variantes : insuffisance de la thyroïde, insuffisance glandulaire, insuffisance thyroïdale, hypothyroïdie.

Le graphe permet de représenter ces variantes en explicitant leur équivalence. La représentation de l'équivalence interviendra dans le processus de traduction que l'on peut voir comme une correspondance entre deux graphes rassemblant les formes synonymes des deux langues concernées.

On notera bien qu'il sera nécessaire de construire des dizaines de milliers de telles représentations afin de localiser dans les textes et de traduire les expressions les plus courantes. Pour les expressions techniques, l'absence de normalisation se traduit par l'existence de variantes terminologiques nombreuses qui sont représentables par la même méthode.

Les composantes lexicales d'un système de traduction qu'on a évoqué ne font que décrire les formes, pas le sens.

- Synonymie et différence de sens :

Si on revient à l'exemple précédent, il est légitime de considérer la forme 'hyperthyroïdie' non pas comme une variante morphologique de l'un des mots composés du graphe, mais comme un synonyme, autrement dit 'hypothyroïdie' est une expression sans lien formel avec les autres. Mais on notera bien que les difficultés pratiques de détermination des formes synonymes sont indépendantes du système de représentation par graphes. Les dictionnaires distinguent les différents sens des mots à l'intérieur des articles de dictionnaire, mais pour associer un seul de ces sens à un mot isolé dans un texte par un ordinateur, il est nécessaire de disposer d'informations contextuelles nombreuses et formalisées. En effet, les bases formelles des distinctions de sens peuvent être localisées dans des composants très variables de la phrase, par exemple :

- Dans une préposition :

Luc pèse sa décision

Luc pèse dans sa décision

- Ou dans un article :

Luc a raison de faire cela

Luc a une certaine raison de faire cela

Mais dans d'autres contextes les mêmes différences de formes n'auront pas d'incidence sur le sens :

Luc habite dans un palais = Luc habite un palais

Luc a peur de faire cela = Luc a une certaine peur de faire cela

Les différents sens d'un mot sont considérés comme autant de mots séparés, ainsi, dans les exemples qui précèdent, nous avons présenté deux verbes 'peser', deux substantifs 'raison', un verbe 'habiter', etc. Cette situation est générale et pour la décrire à l'échelle du lexique français, il est nécessaire de définir un programme de description qui généralise le programme transformationnel de Z.S.Harris. Le programme du lexique-grammaire consiste à décrire chacun des mots de la langue dans une phrase élémentaire et à énumérer toutes les transformations possibles de cette phrase. En général, l'ensemble des phrases où entre un mot, phrases de base ou phrases transformées, est caractéristique du mot. Autrement dit, étant donné deux mots ayant la même forme, on trouve en général des phrases qui les caractérisent. Dans de telles phrases, on observe uniquement l'un des sens. Rien ne garantit que l'emploi qui se fera fait de ces mots dans les textes sera limité à ces phrases, c'est en effet une situation théorique de levée des ambiguïtés. Comme les procédés syntaxiques permettent d'arriver à un tel résultat, on peut espérer que les traces formelles des opérations syntaxiques en jeu permettront à un ordinateur de lever la majorité des ambiguïtés.

Il devrait être clair que les dictionnaires et éléments de grammaire qu'on vient de présenter sont des composants nécessaires de tout système d'analyse automatique des textes et donc de traduction. Si on revient aux études en cours, qu'elles soient théoriques (élaboration de modèles) ou appliquées (recherche de collocations dans les textes ou d'informations dans les dictionnaires de commerce), il est clair qu'elles n'ont guère à voir avec la construction des dictionnaires et grammaires électroniques qui pose de nombreux problèmes théoriques et pratiques.

II.4.Exemple : les dictionnaires du systeme systran

Quelle que soit la complexité des programmes linguistiques qui représentent 130 000 lignes de programmation en moyenne par paire de langues, leur importance est relativement faible par rapport à celle des dictionnaires eux-mêmes.

En effet, le coût des dictionnaires représente près de 90 % du coût de développement d'une paire de langues considérée. S'il est en effet facile de développer des dictionnaires de 20 000 ou 30 000 termes, leur complexité s'accroît très vite avec l'augmentation de la taille, car les descriptions techniques générales ne sont plus suffisantes pour distinguer les différents sens d'un mot donné dans un domaine particulier.

Un dictionnaire classique (papier ou électronique) va préciser les points suivants pour un nom: son orthographe, le genre, le nombre, sa traduction... ce qui représente 5 ou 6 «codes», alors que l'encodage complet dans le dictionnaire du SYSTRAN comporte jusqu'à 165 positions qui permettent de définir le mot ou l'expression le plus complètement possible, afin que le système puisse prendre les options nécessaires le moment voulu.

Les dictionnaires peuvent être classés comme suit :

1) Le dictionnaire des mots de base :

il contient pour chaque mot de la langue source une description de celui-ci à la fois morphologique, grammaticale, syntaxique et sémantique, complétée par la traduction de base dans chacune des langues cibles accompagnée par l'ensemble des indications grammaticales et syntaxiques correspondant à la synthèse.

Dans ce dictionnaire, les différents homographes sont traités en tant que substantif, verbe ou adjectif... cependant les polysémies ne sont pas traitées à ce niveau, mais relèvent des dictionnaires contextuels.

2) Les dictionnaires d'expressions :

C'est à ce niveau que les mots seront traités en fonction du contexte dans lequel ils se trouvent, en répartissant les dictionnaires selon la liste ci-après :

- a) **Dictionnaire idiomatique** dans lequel seront classées les expressions invariables, par exemple «*à la queue leu leu*».
- b) **Dictionnaire des groupes nominaux** : il s'agit là aussi d'expressions figées, mais variables en nombre, du type «*pomme de terre*», «*chemin de fer*», ou «*ville a"eau*».
- c) **Dictionnaires analytiques**: ils traitent des exceptions aux règles grammaticales stockées dans les programmes. Un sous-ensemble, les dictionnaires de résolution homographique, permet de prendre en charge certains termes homographiques en fonction d'un contexte très précis, du type «*rendre compte*», «*tenir compte*», ou «*prendre pied*».
- d) **Dictionnaire contextuel**: celui-ci est de loin le plus complexe, car c'est lui qui doit permettre de faire le choix final entre les différents sens d'un terme. Ainsi le mot «*oil*» en anglais nécessite le codage de plus de 400 contextes pour pouvoir déterminer le sens à retenir parmi les trois possibles : «*huile*» (mécanique ou alimentaire), «*pétrole*», ou «*parfum*». Le terme «dictionnaire» est d'ailleurs ici mal choisi, car il s'agit en fait d'une base de données de microprogrammes linguistiques (syntaxiques et sémantiques) exécutés sur la base d'un mot clé source.

Ainsi peuvent être prises en considération des différences telles que :

- «*Il fait beau*» ou «*il a beau faire*».
- Prendre «*un*» verre ou prendre «*le*» verre.

Ces dictionnaires constituent un système expert organisé dans une base de données en dix niveaux, subdivisés chacun en huit niveaux de priorité. Le tout est géré par un véritable moteur d'inférence indépendant de la langue source et de la langue cible, et qui définit à chaque étape l'applicabilité de chaque règle lexicale.

BIBLIOGRAPHIE

ALSTON, W. P. (1971) : «How Does One Tell Whether a Word Has One, Several or Many Sensés ?», Steinberg, D.D. et L.A. Jakobovits (eds.), *Semantics*, Cambridge, Cambridge University Press, pp. 35-47.

ANDERSON, J. D. (1972) : «The Development of the English-French, French-English Bilingual Dictionary: A Study in Comparative Lexicography», *Word*, 8, pp. 1-144.

COHEN, M. (1962) : «Le fait dictionnaire», *Proceedings of the 9th International Congress of Linguists*, The Hague, pp. 497-503.

CRUSE, D. A. (1986) : *Lexical Semantics*, Cambridge, Cambridge University Press.

DELISLE, J. (1984) : *L'Analyse du discours comme méthode de traduction*, Ottawa, Éditions de l'Université d'Ottawa.

HARTMANN, R. R. K. (1983) : «Introductory Essay», *Proceedings of «Lexeter 83»*, Exeter, Niemeyer. **HOUSEHOLDER, F. W. and S. SAPORTA** (eds.) (1975): *Problems in Lexicography*, Bloomington, Indiana University Publications.

KATZ, J. J. et J. A. FODOR (1963) : «The Structure of a Semantic Theory», *Language*, 39, pp. 170-210. **LAROSE, R.** (1987) : *Théories contemporaines de la traduction*, Sillery, Presses de l'Université du Québec.

LARSON, M., 1999, La traduction signification-meaning-based ; un guide des equivalences en croix-langue.

LEDERER, M. (1973): «La traduction: transcoder ou réexprimer?», *Études de linguistique appliquée*, 12, Paris, Didier.

NEWMARK, P., 1981, *Approaches to Translation*, Pergamon Press, Oxford.

MOSKOWITZ, D. (1973) : «Le traducteur: récepteur et destinataire du message», *Études de linguistique appliquée*, 12, Paris, Didier.

SAUSSURE, F. de (1973) : *Cours de linguistique générale*, édition critique préparée par

Tullio de Mauro, Paris, Payot.

STEINBERG, D. D. and L.A. JAKOBOVITS (eds.) (1971) : *Semantics*, Cambridge, Cambridge University Press.

TATILON, C. (1982) : «Traitement des unités lexicales», *Meta* 27-2.

VAN HOOF, H., 1994, Traduire l'anglais, Duculot.

VINAY, J.-P. et J. DARBELNET (1977) : *Stylistique comparée du français et de l'anglais*, Montréal, Beauchemin.

WEINREICH, U. (1975) : «Lexicographie Définition in Descriptive Semantics», Householder, F. W. and S. Sa-porta (eds.), *Problems in Lexicography*, pp. 25-44.

WTEZELL, R. J. (1975) : «Problems in bilingual lexicography: Romance and English», *Hispania*, 58-1, pp. 131-136.

WIGGTNS, D. (1971): «On Sentence-sense, Word-sense and Différence of Word-sense. Towards a Philosophical Theory of Dictionaries», *Semantics*, Cambridge, Cambridge University Press, pp. 14-34.

Chapitre V : la traduction automatique et la langue arabe

I. Introduction:

Le traitement automatique de la langue arabe n'a pas encore atteint le même niveau de recherche et de résultats que les autres langues (notamment l'Anglais et le Français). Cela n'empêche pas que des tentatives, pas toujours fructueuses, furent menées qu'on va essayer d'en donner un aperçu pour mieux cerner les difficultés qui l'entourent. Ces difficultés découlent presque toutes de la nature de la langue arabe qui est caractérisée par sa forte morphologie

Ainsi, on présente dans ce chapitre une vue globale de la structure générale de la base de donnée lexicales qui est nécessaire pour tout traitement automatique de la langue arabe fondée sur une approche déclarative de la morphologie. Il est bien connu que les entrées lexicales peuvent être distinguées sur la base de caractéristiques morpho-syntactiques qui sont, dans une large mesure, reliées au sens. Ces caractéristiques sont considérées à la lumière des relations entre la grammaire et le lexique. Le but recherché est de donner une définition formelle des entrées en langue arabe d'une base de données lexicales monolingue ou multilingue par rapport au problème crucial de la lemmatisation.

Dans ce sens, l'analyse morphologique de l'arabe a dévoilé beaucoup de problèmes qui demeurent controversés. On abordera, notamment, la structure du verbe en langue arabe et son rapport avec le verbe français et anglais.

I.1.La théorie du mot graphique en arabe :

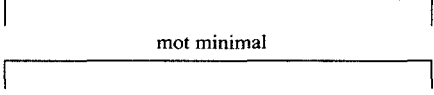
Le *mot graphique en arabe* comporte une structure d'objet complexe. D. Cohen (1961/70) appelait *mot maximal* l'unité décomposable en : *proclitique(s), préfixe, base, suffixe(s), enclitique(s)* — ci-après : *PCL, PRF, BAS, SUF, ECL* (terminologie actualisée ; cf. Desclés *et al.* 1983). On trouvera dans le tableau ci-dessous un exemple simplifié.

La *base*, pour la partie du lexique qui relève du système dérivationnel propre aux langues sémitiques¹⁰, s'analyse en racine (*RAC*) et schème (*SCH*), au sens que prennent ces termes dans le domaine sémitique. On notera toutefois qu'un sous-ensemble important des noms ne peut être analysé ainsi. Ces noms correspondent à des *pro-bases (PBA)*.

Ex. : <YaASiMiYN>, «jasmin» ; <? iBRaAHiYM>, «Abraham», etc.

Bases et pro-bases sont le *noyau lexical* du mot graphique (ou *formant-noyau, Fn*), les autres constituants étant des *extensions (ou formants-extensions, Fe)*.

En faisant usage des *frontières faible* et *forte* de morphème (respectivement «#» et «+»), on peut représenter le mot ainsi :

Représentation «classique» en constituants immédiats Exemple (sommaire) :	mot maximal  ## PCL # PRF + {BAS OU PBA} + SUF # ECL ## ## Li # Ta + KTuB + uW # Hu ## «pour que» pronom 2^e pers. «écrire» inachevé plur. masc. (subjonct.) «lui» pronom complém., accusatif
Représentation faisant apparaître la saillance du noyau lexical	Fn 

Le schéma du mot

■ Pour la représentation en constituants immédiats (Desclés *et al.* 1983), lire «##» comme *frontière de mot* ; le critère empirique permettant de distinguer la frontière «+» (pré- ou suffixation) de la frontière «#» (enclise : pro- ou enclitiques) est celui de la pause potentielle (J. Lyons) : en l'absence du PRF ou du SUF auquel elle est liée par une frontière «+», la BAS — ou la PBA — ne peut constituer une forme libre minimale (Bloomfield). En revanche, elle peut, de ce point de vue, «se passer» des ECL et des PCL. Pour une définition de chacun des termes cités dans ce paragraphe, voir Dichy & Hassoun (1989 «Lexique de définitions» : 265-276) et Dichy (1990 : ch. X).

I.1. 1.Relations contextuelles et formants de mots :

La différence entre les deux représentations ci-dessus est

considérable. Dans le modèle de D. Cohen (1961/71), les relations entre les morphèmes inclus dans le mot demeuraient conçues sur un mode distributionnel et concaténatoire.

La représentation introduite dans Desclés *et al.* (1983) et développée dans Hassoun (1987), Dichy & Hassoun (1989), ajoutait notamment l'idée fondamentale d'un dictionnaire des bases associé aux traitements en synthèse et en analyse, la distinction entre ces deux modes de traitement et la prise en compte, dans cette double perspective, des processus morphophonologiques de transformation.

La représentation faisant apparaître la saillance du noyau lexical constitue un nouveau saut qualitatif: le Fn est un formant de mot, *i.e.* un formant dont les relations contextuelles sont spécifiées dans le domaine de cette unité. Il s'agit dès lors de décrire le système de connaissances correspondant à l'ensemble des relations entre les unités du dictionnaire associé aux traitements du niveau du mot. Au sein de cette unité, les relations sont, au niveau le plus abstrait, de deux types :

(a) <Fe - Fe> et (b) <Fn - Fe>

Le second (b) constitue le champ des spécificateurs du noyau lexical, ce qui ouvre la voie à la définition de l'unité lexicale en arabe présentée ci-dessous.

I.1.2. Le schéma de la grammaire des formants du mot :

Une *grammaire des formants du mot* issue de cette conception comporte

deux sortes de *relations contextuelles* (Dichy 1987) :

■ la *relation d'ordre* au sens strict régit la position respective des formants, qui, dans le cadre du mot, est strictement prédictible. On dira que cette unité peut être représentée au moyen d'un vecteur ordonné.

Ex. : La préposition proclitique <Li#>, «pour» est placée avant l'article, et après les coordonnants <Wa#> ou <Fa#> ;

■ les *relations de collocation morphophonologiques et syntaxiques* «gèrent» les incompatibilités et cooccurrences, ainsi que les modifications morphophonologiques entraînées par la présence d'un formant.

Ex.: L'article <?aL#> est incompatible avec les marques casuelles de l'indétermination, mais aussi avec un très grand nombre de noms propres. Le pronom clitique <#Hu> est réalisé <#Hi> après une voyelle <i>.

I.1.3. L'évolution des modèles de représentation des phénomènes du niveau du mot en arabe :

Si l'on demeure ici dans un modèle «item et agencement» (comme dans les démarches computationnelles liées à la morphologie à deux niveaux — Fradin 1994), la différence avec le modèle de D. Cohen et ceux qui en découlent plus ou moins directement réside essentiellement dans la

représentation des systèmes de connaissances relatifs au mot graphique dans une perspective d'intelligence artificielle (Dichy 1993) : traitement déclaratif d'ensembles complexes de données associées à chacune des unités du lexique (Hassoun 1987), bases de données relationnelles soumises à la contrainte générale de compatibilité avec la reconnaissance et la génération (Dichy 1990).

La différence ne se situe pas seulement au niveau de la structure générale des systèmes de traitement envisagés, elle comporte également des aspects descriptifs (en linguistique) et de pure faisabilité (en linguistique informatique). Les modèles issus de celui de D. Cohen ont tous le mérite de souligner l'importance de ce qui correspond, ici, au schéma des relations <Fe — Fe>. Mais ils ne prennent pas en compte les relations <Fn — Fe>, c'est-à-dire le jeu des spécificateurs morphosyntaxiques associés aux Fn, sans lesquels il est exclu de parvenir à éliminer la production d'un grand nombre de formes exclues de la langue en génération, ou un niveau très élevé de «bruits» en analyse.

Ex. : <?aNna>, «geindre», est intransitif. En mode synthèse, *<?aNna#Hu>, «geindre-lui», où <#Hu> serait un pronom proclitique complément d'objet, doit donc pouvoir être exclu (au moyen du M-spécificateur «verbe intransitif») ; en mode analyse, seule devra être retenue la forme homonyme <? aNna>, conjonction «que», suivie du pronom <#Hu>, «il» ou «lui».

À cela s'ajoute, dans le cadre du modèle SAMIA, la possibilité de produire, en génération, des formes en graphie vocalisée et de reconnaître, en

mode analyse, des mots graphiques non vocalisés.

I.1.4. Structure de l'unité lexicale simple (UL) en arabe :

En théorie de l'écriture, on peut interpréter la structure <Fe — Fn — Fe> du mot graphique en arabe comme un reflet de sa fonction logographique : le mot est en effet l'unité perceptive susceptible d'inclure un noyau lexical (nom, verbe) *au plus*. Il n'existe pas en arabe, à un petit nombre d'exceptions près, de composition de radicaux (au sens où l'on entend ce terme dans les langues indo-européennes), comme dans *métro+pole*, *théo+sophie*, etc. : la composition doit recourir à des procédés syntaxiques, *i.e.* à des phrasèmes.

Tout se passe comme si le système d'écriture constituait les mots incluant un noyau lexical en unités ayant pour fonction d'en manifester la présence : toutes les relations associées aux *formants-extensions* (Fe) de cette unité complexe convergent, en compréhension, vers la détermination de la catégorie syntaxico-sémantique du *formant-noyau* (Fn). Cette analyse «en faisceaux» des textes en écriture arabe non vocalisée n'est possible qu'en raison de la présence d'informations relatives aux relations de collocation morphophonologiques et syntaxiques associées aux signifiant, signifié et syntactique des Fn dans la mémoire lexicale des sujets (Dichy 1990 : ch. IX et X) — informations représentées, dans le modèle de traitement automatique auquel il est fait ici référence, par un jeu de spécificateurs morphosyntaxiques régissant les relations de collocation entre le Fn et les Fe.

Or, une précision essentielle doit être apportée : en ce qui concerne les noms (et à la différence des verbes), l'unité lexicale simple ne coïncide pas

toujours avec le *formant-noyau* (Fn). Certains *formants-extensions* (Fe) sont en effet susceptibles, lorsqu'ils sont associés à une base nominale, de se trouver, pour ainsi dire, pris avec elle dans un processus de lexicalisation. Un formant-extension sera dit lexicalisé (appelé *formant-extension lexicalisé* — *Fel*) lorsque l'unité <Fn,Fel> résultant de son association à un formant-noyau donné constitue une unité du lexique indépendante. Une *unité lexicale simple* (UL) est donc constituée :

■ soit d'un Fn (on écrira : UL = < Fn >) ; c'est le cas de toutes les bases verbales et des bases nominales dépourvues de Fe lexicalisé (Fel).

Ex. : < MaKTaB > (plur. < MaKaATiB >) «bureau», où l'on a : UL = < Fn = MaKTaB > ;

n soit d'un ensemble UL = <Fn,Fel>, où Fel peut inclure plus d'un formant, et dont l'ordre séquentiel est assigné par la grammaire des formants du mot (*relation d'ordre* au sens strict).

Ex.: <MaKTaBa&> (plur. <MaKTaBaAT>), «bibliothèque» ou «librairie», constitue une UL distincte de la précédente, analysable ainsi : UL = <Fn = MaKTaB, Fel = + a& >. Dans le nom propre < ? aL#TaAHiR >, l'article d'«excellence» < ? aL#> est un Fel.

I.1.4.1. Le fléchage entre unités lexicales :

Un deuxième trait définitoire vient s'ajouter au précédent: celui du

fléchage entre unités lexicales, que l'on peut représenter sous la forme : UL]
<—> UL_n (où UL_n = UL₂, UL₃, UL₄...).

Ce trait reflète, au départ, une propriété de la morphologie, notamment, des langues sémitiques : le passage d'une forme nominale au singulier à son correspondant au pluriel (dans un nombre important de cas, il y a plusieurs pluriels), ou de la forme du verbe à l'accompli à celle de l'inaccompli, du verbe au nom de procès (*masdar*), etc. Ce passage peut avoir lieu, selon la terminologie traditionnelle :

■ soit par «*dérivation externe*», i.e. par suffixation ;

Ex.: <MuDaRriS>, «enseignant» a pour pluriel masculin: <MuDaRriS + uWNa> (pluriel construit avec le suffixe < + uWNa >). Le participe actif < KaATiB >, «en train d'écrire» ou «ayant écrit» a de même un plur. masc. < KaATiB+uWNa > ;

■ soit par «*dérivation interne*», i.e. par une modification du schème de la base, la racine demeurant inchangée.

Ex. : UL_j = <HaDiYT>, «moderne», de racine =HDT et de schème <R'aR²iYR³> (tel que R¹ = H, R² = D et R³ = T), a pour pluriels UL₂ = < HuDalaA? > et UL₃ = < HiDaAl >, de même racine, mais de schèmes différents, respectivement, <R'uR²aR³aA? >, et ^R'iR^AR³^.

On notera que dans les cas de polysémie ou, plus rarement, d'homonymie de la forme correspondant à l'entrée-vedette du dictionnaire, des pluriels

distincts sont susceptibles d'être associés à des sens différents : on aura alors, dans la base de données lexicale, des UL distinctes. (Le même phénomène est observable pour la relation entre les verbes de la forme simple et les noms de procès — *masdar-s.*)

Ex. : En français, on a, pour *ciel*, deux pluriels de sens en partie différents : *deux* et *ciels*. En arabe, l'unité lexicale UL'i = <HaDiYT>, «discours rapporté, récit, conversation (...)», est reliée par fléchage aux pluriels UL'₂ = < ? aHaADiYT> et UL'₃ = < HiDTaAN >. Cette unité doit être distinguée de l'UL[= < HaDiYT > présentée dans l'exemple précédent.

On dira qu'il y a *fléchage* entre deux UL (ou plus) lorsque l'une d'entre elles est supplétive de l'autre, c'est-à-dire lorsque l'une d'entre elles remplace l'autre dans les paradigmes morphosyntaxiques d'une manière qui ne peut être prédite au sens strict par des règles opérant en génération, ou qui présente un caractère d'opacité en reconnaissance.

Ex. : Ainsi, comme on vient de le voir avec l'exemple de < HaDiYT >, pour un même schème de singulier <R'aR²iYR³> (au demeurant fort fréquent), plusieurs schèmes de pluriels coexistent dans la langue. On ne peut considérer les fléchages que l'on vient de rencontrer comme prédictibles, ni en mode synthèse ni en mode analyse. (Il faudrait pour cela que l'on puisse constater une relation bijective entre schèmes du singulier et du pluriel, ce qui ne se vérifie quasiment jamais en arabe sur l'ensemble d'une série'².)

Cette définition donne au concept de fléchage une extension plus large que celle de la «dérivation interne». Deux types de cas, non compris dans cette dernière, sont en effet inclus :

■ le fléchage entre une unité lexicale coïncidant avec le formant-noyau, et une UL

comportant un noyau et un formant-extension lexicalisé, *i.e.* le fléchage de type :

$UL_1 = \langle Fn \rangle \longleftrightarrow UL_2 = \langle Fn, Fel \rangle$ ou, inversement : $UL_1 = \langle Fn, Fel \rangle \longleftrightarrow UL_2 = \langle Fn \rangle$.

Ex.: $\langle KaATiB \rangle$, «secrétaire», «écrivain» ($UL = \langle Fn \rangle$) a pour l'un de ses deux pluriels: $\langle KaTaB+a\& \rangle$, où $F_n = \langle KaTaB \rangle$ et $F_{el} = \langle +a\& \rangle$. Inversement, on a: $\langle HaQiYB+a\& \rangle$, «bagage», «valise», ($F_n = \langle HaQiYB \rangle$; $F_{el} = \langle +a\& \rangle$); pluriel $\langle HaQaA? iB \rangle$, de structure $UL = \langle Fn \rangle$;

■ le fléchage entre bases de racines différentes.

Ex.: $\langle ?iMRa? + a\& \rangle$, «femme», de racine $=MR?$, a pour pluriel $\langle NiSaA? \rangle$, de racine $=NSW$ (le fléchage met en relation un F_n relevant d'une racine R_1 avec un F_{n2} , de racine R_2). Ce phénomène de supplétion est par ailleurs bien connu : cf. en français : *je vais* $\longleftrightarrow$ *nous allons*, ou des «paires» telles que *cerf* $\longleftrightarrow$ *biche* ; *lièvre (ou bouquin)* $\longleftrightarrow$ *hase*.

Les relations de fléchage entre unités lexicales telles qu'elles se dégagent du jeu de critères formels dont le schéma est présenté ici fera, bien entendu,

l'objet d'une étude séparée.

I.1.4.2. Fléchages et relations contextuelles :

Fléchages et relations contextuelles jouent, au sein de la partie de la grammaire de formants traitant de l'insertion de l'unité lexicale dans le mot, des rôles complémentaires. Le pluriel d'une UL nominale sera ainsi constitué :

■ soit de son Fn du singulier, suivi des suffixes du pluriel : le pluriel est alors régi par les *relations contextuelles*.

Ex. : < SaAKiN >, «qui habite» (participe «actif» ; cf. en français, le participe présent «habitant»), pluriel : < SaAKiN + uWN > (avec le suffixe du masc. < +uWN >) ;

ii soit d'un Fn de même racine, mais de schème différent (avec ou sans Fel associé) : le pluriel est, ici, obtenu par *fléchage*.

Ex. : La même forme <SaAKiN>, quand elle devient un nom à part entière (*i.e.* lorsqu'elle cesse d'être un déverbal), a pour pluriels <SuKKaAN> et <SaKaN+a&> lorsqu'elle prend le sens d'«habitants» (avec -s; cf. également le pluriel nominal *résidents*, la forme participiale étant orthographiée -ant), et <SaWaAKiN >, lorsqu'elle signifie «segment-lettre quiescent»¹³.

I.2. Les spécificateurs morphosyntaxiques du domaine du mot :

I.2.1. Les deux types de M-spécificateurs :

Les spécificateurs morphosyntaxiques du domaine du mot (M-spécificateurs) associés aux bases nominales et verbales relèvent en conséquence de deux grands types :

- l'un «horizontal» (de nature syntagmatique), correspond à l'ensemble des relations de collocation morphophonologiques et morphosyntaxiques entre le formant-noyau lexical et les formants-extensions (par ex., les verbes intransitifs sont incompatibles avec les pronoms enclitiques correspondant à des compléments d'objet) ;
- l'autre «vertical» (de nature paradigmaticque), est celui des *fléchages*, eux-mêmes répartis en deux catégories : (a) $UL_j \longleftrightarrow UL_n$, comme ci-dessus, mais aussi : (b) $BAS_j \longleftrightarrow BAS_n$, *i.e.* entre bases relevant de racines soumises à des transformations (voir ci-dessous, S-4.1).

I.2.2. Le schéma des spécificateurs morphosyntaxiques du domaine du mot :

Considérons maintenant la liste fermée des *M-spécificateurs*. Je ne la reproduirai pas ici intégralement, me contentant d'en exposer le schéma et d'analyser ce qui relève en propre de la production et/ou de la reconnaissance. On peut considérer cinq grands types, numérotés ci-dessous de S-0 à S-4 («S» pour «spécificateur»).

S-0 - La catégorie lexicale majeure : bases de noyau lexical soit nominal, soit verbal. (Dans le domaine du mot, cette répartition, par ailleurs traditionnelle dans les sciences médiévales arabes du langage, est

suffisante.)

S-1 - La structure de l'unité lexicale, qui peut être, comme indiqué, de la catégorie soit $UL = \langle Fn \rangle$, soit $UL = \langle Fn, Fel \rangle$. Ce spécificateur n'affecte que les bases de noyau nominal (les UL verbales étant toujours du type $UL = \langle Fn \rangle$).

S-2 - Des variables affectant la catégorie lexicale majeure, pertinentes pour le traitement du niveau du mot : le paradigme de conjugaison de la base verbale concernée (achevé, inachevé dit «indicatif», «subjunctif» ou «apocope», impératif, etc.) ; la catégorie de la base nominale, en termes de : sing. (masc, fém.) ; pluriel (masc, fém.) ; collectif; nom propre, nom commun...

S-3 - Les relations de collocation, elles-mêmes réparties en deux sous-catégories :

- les collocations faisant intervenir des formants-extensions lexicalisés (Fel) ;

- les collocations corrélées à la structure morpho- ou sémantico-syntaxique de l'UL. Citons :

- Bases verbales (formulation abrégée de : «bases à noyau lexical verbal») :

- a) Le «numéro de module» identifié par Abu Al-Chay (1988), ou par Ammar et Dichy, qui désigne le type du verbe en fonction de son modèle de conjugaison.

b) Le caractère transitif (T) ou non (nT) du verbe, c'est-à-dire sa compatibilité, ou son incompatibilité avec un pronom complément (nécessairement, en arabe, enclitique).

c) La règle générale qui interdit en arabe la cooccurrence d'un complément d'objet de la première ou de la deuxième personne et d'un sujet à la même personne, doit

être levée avec un petit nombre de verbes (traditionnellement: *?afa:l al-qulu:b* et *?afad al-hiss*, «verbes de sentiment» et «de perception»).

Ex. : $\langle ?a + DuNn + u\#NiY \rangle$ («je me crois») ; $\langle Ta + RaA\#Ka \rangle$ («tu te vois», au sens «mental» du verbe voir).

• Bases nominales (formulation abrégée de : «bases à noyau lexical nominal») :

■ La liste des suffixes (SUF) de cas compatibles avec la base considérée. Cette catégorie de spécificateurs prend en charge les déclinaisons dites «diptotes», etc.

S-4 - Les fléchages, qui ont été présentés ci-dessus. Ils peuvent être répartis en deux types généraux.

S-4.1 - Le premier concerne les modifications phonologiques affectant certaines bases en fonction des suffixes ou préfixes qui leur sont associés. Il s'agit du fléchage :

$BAS] \longleftrightarrow BAS_n$.

♦ Bases verbales :

Les paradigmes de conjugaison des verbes de racines anomaux comportent plusieurs réalisations de la base verbale, chaque réalisation étant compatible avec un sous-ensemble déterminé de la liste des suffixes.

Ex. : Le verbe <MaDda>, «tendre», «étendre», comporte, à l'achevé, deux bases : $BAS_1 = \langle MaDaD + \rangle$ (en graphie non-vocalisée : <MDD>), comme dans <MaDaD+Tu>, < MaDaD + Na >, etc., (suffixes ne commençant pas par une voyelle) ; $BAS_2 = S MaDd + \rangle$ (en graphie non vocalisée : < MD >), comme dans : < MaDd + aT >, <MaDd + uW >, etc., (suffixes commençant par une voyelle).

Chacune de ces bases anomaux doit être enregistrée dans le dictionnaire du système, avec l'indication de la liste des suffixes (du «vecteur-suffixe») avec laquelle elle est compatible. L'enregistrement des BAS_2 , BAS_3 , etc., se fait par génération automatique à partir de la seule saisie de la forme qui désigne verbe dans la grammaire traditionnelle. Une telle démarche est due aux grandes difficultés que l'on rencontre lorsque l'on cherche à écrire des règles opérant en reconnaissance, à partir des règles phonologiques, dont le formalisme concorde parfaitement avec la simulation de processus opérant en production : il faut, en effet, clairement distinguer entre la prédictibilité par les règles en mode synthèse et le caractère reconnaissable en mode

analyse (Dichy 1995).

» **Bases nominales** :

Il s'agit de la modification susceptible d'affecter les bases nominales associées au suffixes < + iYY > du nom-adjectif de relation. Celle-ci n'est prédictible par des règles (au sens strict) que pour un sous-ensemble du lexique, les grammaires scolaires présentant de longues listes d'exceptions.

Ex. : Dans < MaDiYN + a& >, «ville» et < MaDaN + iYY >, «urbain», la règle de modification du schème s'applique; dans : < TaBiY^c + a& >, «nature» et < TaBiY^c + iYY >, «naturel», par ailleurs de schème identique, la règle ne s'applique pas. (On peut démontrer que ce phénomène ne dépend nullement de la structure phonologique des noyaux lexicaux.)

S-4.2 — Le deuxième type de fléchage, dont nous avons donné des exemples lors de la présentation de cette notion (§ 2.4.1.), est le fléchage $UL_j \longleftrightarrow UL_n$.

• **Bases verbales** : On a principalement :

- la relation entre *l'achevé* et *l'inachevé* de la forme simple (non augmentée) du verbe (problème de la voyelle de la deuxième radicale) ;
- la relation entre le verbe et ses dérivés nominaux immédiats, notamment : le nom de procès (*masdar*), dont la forme n'est pas prédictible pour le verbe «non augmenté», ainsi que les participes «actif»

{? *ism al-fa:^cit*) et «passif» {? *ism al-mafu.J*).

Ex. : Pour le verbe < HaZiNa> «être triste», le lexique n'atteste pas le participe «actif» *<HaAZiN>, «attristé». On a en revanche une forme adjectivale (*sifa musabbaha*) <HaZiYN>. Pour le verbe de sens voisin <Ba?iSa>, «être désespéré», la forme adjectivale attestée < BaA ? iS > («désespéré»), est du même schème < R¹aAR²iR³ > que le participe actif.

Ces exemples montrent que le lexique doit contenir les informations permettant :

- (1)d'attester l'existence d'un participe actif ou d'une forme adjectivale, et
- (2)d'indiquer, pour la forme attestée, le schème correspondant.

Si (1) peut être déduit de la valeur sémantique du verbe, (2) relève au sens strict du syntactique, tel qu'il a été défini par Mel'cuk. Les spécificateurs morphosyntaxiques portant sur le point (1) suppléent, comme indiqué (§ 1.4.1.), l'absence de signifiant — au sens que prend ce terme chez l'agent cdgnitif humain — dans une base de données informatisées.

« Bases nominales :

Il s'agit principalement de relations relevant, comme on l'a vu, de la «dérivation interne», telles que : sing. <—> plur., masc. <—> fém., collectif <—> singulatif, etc., auxquelles il faut ajouter des fléchages faisant intervenir deux racines.

I.3. Conclusion : le caractère fini des m-spécificateurs et l'inventaire exhaustif des données du niveau du mot :

Un postulat sous-tend, généralement sur le mode implicite, toute entreprise lexicomatique : celui de la projection de traits en inventaires finis sur les unités en inventaire non fini du lexique. Les entrées d'une base de données lexicales étant de structure <chaîne de caractères, spécificateurs>, le principal problème est, comme indiqué, de dresser un inventaire de spécificateurs qui soit à la fois *fini* et *exhaustif* dans un domaine d'extension contextuel donné.

Par *exhaustif* il faut entendre : (1°) satisfaisant à la principale exigence de la génération (classiquement: engendrer l'ensemble des formes correctes et elles seules), et (2°) qui ne produise, en reconnaissance, que les ambiguïtés permises par la langue (ambiguïtés morphosyntaxiques de construction) dans le domaine d'extension considéré. Dans Dichy (1993 : 70-71), il a été montré que le nombre d'ambiguïtés permises par la langue dans un domaine donné dépend, d'une part, de l'extension de ce dernier (un domaine étroit produisant plus d'ambiguïtés qu'un domaine étendu), et d'autre part, de la richesse du système de connaissances de l'agent cognitif (humain ou informatique) qui procède à la reconnaissance des données linguistiques. En d'autres termes, pour l'analyse automatique du mot, plus le système est capable d'attribuer à une même entrée de valeurs attestables dans la langue concernée, plus il peut être considéré comme performant. Il en ira de même au niveau de la phrase, étant entendu qu'une phrase donnée comporte en principe moins

d'interprétations différentes possibles qu'un mot (notamment un mot arabe non vocalisé...).

Or, au schéma des spécificateurs morphosyntaxiques du domaine du mot graphique (M-spécificateurs) présenté ici correspond un inventaire à la fois *exhaustif* et *fini*. Cela a été rendu possible par le raisonnement de type ensembliste suivant : le mot graphique comprend en arabe deux catégories d'unités, respectivement des formants-noyaux (Fn) et des formants-extensions (Fe), les premières d'inventaires non finis, les secondes d'inventaires finis ; or la structure perceptive de cette unité (§ 2.4.) la contraint à ne comporter qu'une unité lexicale (un Fn) au plus; il convient donc d'envisager les M-spécificateurs comme la «projection» de traits en nombre fini, parce que «gérant» un ensemble de relations contextuelles associées à des unités d'inventaires finis (les Fe du mot), sur chacune des unités en nombre non fini du lexique, c'est-à-dire sur chacun des Fn de l'arabe (lexèmes composés et phrasèmes provisoirement exclus).

Ce raisonnement découle de la structure de l'unité-mot en arabe, qui vérifie le schéma de contextualisation <morphèmes lexicaux — morphèmes grammaticaux>. L'étape suivante est celle du domaine de la phrase (synthèse de l'arabe écrit vocalisé, et analyse de l'écrit non vocalisé). Les problèmes y sont d'une autre échelle : s'ajouteront à ce schéma des relations <morphèmes lexicaux — morphèmes lexicaux>. Il faudra donc, si l'on veut parvenir à une analyse syntaxique et au passage de phrases arabes en écriture non vocalisée, envisager des formalismes relevant de la famille des grammaires d'unification

(notamment des grammaires affixales étendues). Mais d'ores et déjà, la base de données lexicale réalisée permet, pour la première fois, une reconnaissance de mots écrits non vocalisés.

II. la T.A et le verbe arabe :

II.1. Introduction :

Du fait même de leur simplicité et de leur réduction, les formes verbales arabes posent des problèmes quant à leur informatisation et — par la suite — quant à leur traduction en formes équivalentes dans d'autres langues. Déjà, la structure même du mot arabe, construit à base de racines consonantiques (trilitères ou quadrilitères) et infixes vocaliques (ces derniers ne sont pas reproduits au niveau graphique), posait et pose encore des difficultés quant à son éventuelle informatisation : faudra-t-il partir de la racine consonantique ou du mot voyelle pour former les classes de mots ? Et puisque, à présent, ce qui nous intéresse c'est le système verbal et sa traduction en français dans l'optique de la mise en place d'un logiciel prévu à cet effet, nous concentrerons notre travail sur la correspondance des formes verbales de ces deux langues.

II.2. Problématique de la structure verbale arabe :

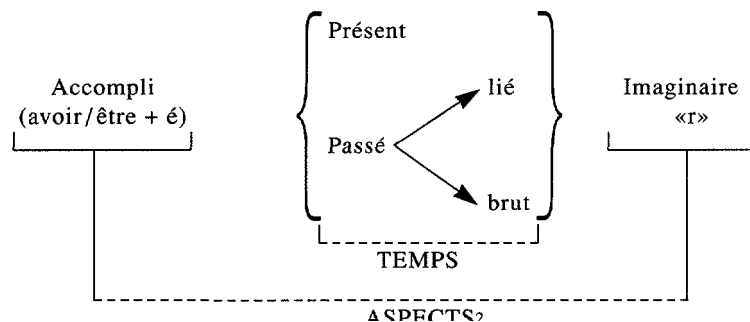
La linguistique descriptive et fonctionnelle avait depuis longtemps souligné la composante aspectuelle du système verbal arabe, lequel ne reconnaît que deux formes morphologiques dont les schèmes sont *fa' al* (*°al-madi* = l'accompli) et *yaf'al* (*°al-mudari* - l'inaccompli). À cette base

bi-aspectuelle viennent s'ajouter des éléments extérieurs pour constituer l'éventail temporel et modal de l'arabe¹.

L'arabe est resté à ce stade de la description de son système verbal, et aucune évolution n'a été signalée depuis le fameux *Kitab* de Sibawayhi (mort au VIII^e siècle = II^e siècle de l'hégire). Or, pour que s'opère un passage de l'arabe au français et inversement, certains obstacles sont à surmonter.

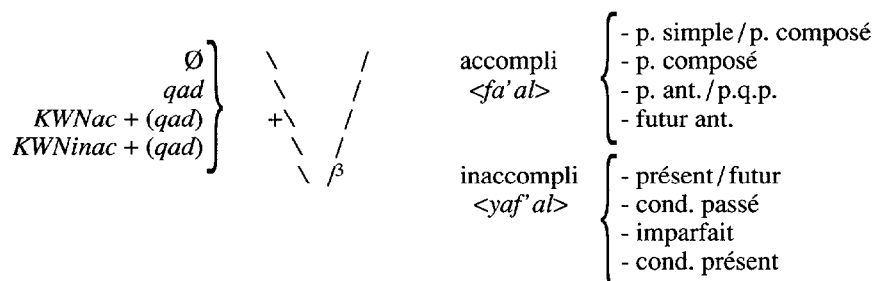
Le premier de ces obstacles émane du fait que, tandis qu'en français c'est le paradigme temporel (présent/passé) qui constitue le pivot stable, obligatoire et exclusif de la distribution verbale auquel viennent s'ajouter les morphèmes aspectuels (avoir/être + é = l'accompli) facultatifs et discontinus, en arabe ce sont les aspects (accompli/inaccompli) qui constituent l'axe paradigmatique stable, et c'est l'expression du temps qui y est facultative. Ainsi apparaît une opposition incontournable quant à la position et à la distribution des éléments obligatoires et facultatifs dans chacune des deux langues en question.

Pour être plus clair, reprenons la règle de réécriture des formes verbales françaises telle que représentée par ce schéma qu'ont élaboré les générativistes français :



En français, c'est l'introduction du morphème de l'accompli par un segment discontinu «avoir/être + é» («avoir/être» se place en tête du SV et «é» à la fin du premier élément conjugable) dans le syntagme verbal qui crée les formes verbales à valeur accomplie. Nous avons ainsi un présent accompli (passé composé = Prés. + <avoir/être + é> + V), un passé lié accompli (P.Q.P. = «ais» + <avoir/être + é> + V), un passé brut accompli (Passé antérieur = «ai» + <avoir + é> + V), un futur accompli (Prés. + «r» + <avoir/être + é> + V) et un conditionnel ou futur du passé accompli («ais» + «r» + <avoir/être + é> + V).

En arabe, la distribution est toute différente. Le caractère accompli ou non accompli de l'expression verbale est déjà donné par la forme brute du verbe dans les schèmes *fa'al* et *yaf'al* (et non par l'introduction d'un morphème équivalent au «avoir/être + é» du français), et c'est l'expression du temps qui se réalise par des éléments surajoutés, comme dans le schéma représentatif suivant :



Un second obstacle se profile à travers la non-équivalence totale entre l'arabe et le français quant à l'expression du passé brut (passé simple) et du passé lié (imparfait) représentant respectivement renonciation historique et renonciation discursive. De désinences différentes en français (ai/ais), ces deux formes du passé font partie du paradigme stable des temps. En arabe, les choses sont différentes. Si le passé brut trouve son équivalent dans la forme fixe de l'aspect verbal accompli (*fa'al*), le passé lié, quant à lui, révèle une construction composée : c'est l'accompli de l'auxiliaire *KWN* (= *kana*) qui vient s'ajouter à la forme aspectuelle inaccomplie du verbe proprement dit (*yaf'al*). Il faudra mentionner à cet effet l'inexistence de la forme participiale en arabe : les deux constituants, auxiliaire et verbe, reçoivent une conjugaison qui, tout en étant différente dans certains cas (auxiliaire à l'accompli + verbe à l'inaccompli = imparfait), n'en est pas moins analogue.

Sur un autre plan, l'unicité de forme que requiert l'aspect accompli (*fa'al*) rend impossible la distinction entre un passé brut historique et un accompli du présent lié au locuteur (respectivement le passé simple et le passé

composé). Nous avons ici une situation inversée par rapport à la précédente. Tandis que l'imparfait arabe requiert un auxiliaire, le passé composé a la même forme brute que le passé simple. Aucune forme morphologique ne venant spécifier le passé composé, on s'en remet au sens de la phrase pour opérer la distinction⁴. Et souvent, l'interprétation de l'accompli *fa'al* est laissée au libre arbitre du traducteur, liberté entraînant parfois quelques anachronismes.

Enfin, la grammaire arabe ne reconnaît pas l'auxiliaire, ou tout au moins ne tranche pas sur la possibilité auxiliante ou copulative de certains verbes appelés incomplets (*°af'al naqisa*)⁵. «Incomplet» signifie que ces verbes ont besoin d'un complément pour parfaire le sens de la phrase. Nous avons là un cas analogue à celui du verbe copulatif français qui ne peut se suffire de son sujet et a besoin d'un autre syntagme (nominal, adjectival ou prépositionnel) pour constituer une phrase. Ainsi, un verbe incomplet comme *KWN* (accompli = *kana* ; inaccompli = *yakunu*) est analysé comme élément verbal s'introduisant dans une phrase nominale et faisant du premier son sujet/agent (*°ismuhu*) et du second son prédicat (*habaruhu*). Cette solution, bien qu'elle soit arrangeante au niveau de surface, n'est pas satisfaisante pour peu qu'on veuille dépasser cette simple constatation superficielle et repenser la notion même de phrase nominale dans sa structure réduite à deux composantes nominales (*°ism* = sujet/agent, et *habar* = prédicat) ou élargie par l'introduction d'un verbe dans l'une des deux composantes.

En effet, une phrase nominale simple comme *°al-waladu maridun*

(l'enfant <est> malade) est toujours interprétée — et partant traduite — comme un état se déroulant dans un présent continu. Avec l'introduction de *kana* (forme accomplie de la racine *KWN*), l'ensemble de l'état est renvoyé dans un passé continu et non-accompli et trouve sa traduction dans l'imparfait français. Cependant, on sait que l'équivalent de l'imparfait (ou de ce qu'on peut désigner par le passé continu et inaccompli) trouve sa réalisation en arabe dans la rencontre de deux verbes, l'un auxiliaire à l'accompli (*kana*) et l'autre à l'inaccompli (*yaf'al*). Cela nous permet de supposer la présence implicite et sous-entendue d'un verbe à l'inaccompli (*yaf'al*) dans la prétendue phrase nominale de surface. Cette supposition est d'autant plus vraie que dans une phrase nominale étoffée comme *kana °al-waladu yadhabu °ila °al-madrasati* (l'enfant allait à l'école) — que l'analyse grammaticale arabe divise en trois parties : *Kana* + *°al-waladu* (= sujet/agent de *Kana*) + *yadhabu °ila °al-madrasati* (proposition ayant la fonction de prédicat de *kana*) —, nous avons la présence explicite de deux verbes, *Kana* + *yadhabu*, où la rencontre de l'accompli de *KWN* avec l'inaccompli du verbe *DHB* donne le passé continu inaccompli, soit l'imparfait.

Cette reformulation de *KWN* comme auxiliaire ou copule amène à une remise en question de deux notions : 1) la fonction de *KWN* comme élément introducteur dans une phrase nominale se déplace vers le syntagme verbal (que ce dernier soit explicite ou implicite), il est donc un élément auxiliaire du verbe ; 2) la notion de phrase nominale est à repenser comme phénomène de surface, et de ce fait elle serait plutôt le résultat d'une transformation

d'effacement du verbe, lequel est nécessairement présent dans la structure profonde.

II.3. Conclusion :

Pour conclure, la démarche que nous venons d'exposer redécouvre dans la structure verbale arabe, à un niveau profond, le système temporel propre à cette langue, système permettant l'élaboration d'une classification unifiée que les informaticiens pourront repenser et redéfinir d'une manière plus technique pour la mise en place d'un logiciel de traduction assistée par ordinateur.

III. Exemples des difficultés de traduction automatique des verbes anglais vers l'arabe :

III.1.Introduction :

Ce qui a été avancé précédemment, concernant l'incompatibilité dans le rapport temporel entre l'arabe et le Français, peut-être projeté à celui de l'Anglais et l'Arabe.

- III.1.1. Les temps du verbe dans la langue anglaise :

La langue anglaise dispose de 12 temps différents pour décrire le comportement verbale caractéristique de l'action :

- 1- ' Présent simple' : est utilisé lorsque l'action se répète régulièrement, pour une vérité général ou une action inscrite dans un emploi du temps.

- 2- 'Présent continuous' : est utilisé lorsque l'action se déroule au moment où l'on parle ou qui continue à l'instant de la parole ou lors d'une action future programmée.
- 3- 'Présent perfect' : lorsque l'action se déroule et s'écoule dans le passé.
- 4- 'Past simple' : lorsque l'action se déroule à un moment précis dans le passé (par le biais d'un adverbe de temps).
- 5- 'Past continuous' : lorsque l'action dure un certain temps dans le passé.
- 6- 'Present perfect continuous' : lorsque l'action démarre dans le passé et dont l'effet continue au moment où l'on parle ou vient juste de se terminer.
- 7- 'Future tense' : l'action est prévue dans le futur.
- 8- 'Future continuous' : lorsque l'action est prévue dans le futur et dont l'effet continue dans le futur.
- 9- 'Past perfect simple' : lorsque l'action s'écoule dans le passé par rapport à une autre action passée explicite qui lui est antérieure.
- 10- 'Past perfect continuous' : lorsque l'action s'entame dans le passé et se poursuit après une action passée citée et explicite.
- 11- 'Future perfect simple' : lorsque l'action va cesser dans le futur.
- 12- 'Future perfect continuous' : lorsque l'action a cessé.

III.1.2.Les temps du verbe en langue arabe :

La langue arabe s'articule sur seulement deux temps :

- L'accompli : lorsque l'action se déroule dans le passé. S'il est précédé de ' قد ', il indique que l'action se déroule dans le passé proche du présent. Par contre il indique le futur lorsqu'il invoque une invitation ou synonyme d'une condition.
- L'inaccompli : il indique que l'action se déroule dans le temps présent ou dans le futur selon les situations :
 - 1- S'il est précédé de prépositions, il indique seulement le futur :
 - 'س' : indique que l'action est imminente ou va se produire dans un futur proche.
سيحل الظلام
 - 'سوف' : indique que l'action va se dérouler dans un futur lointain.
سوف يأتي
 - 2- L'inaccompli peut indiquer l'accompli s'il est précédé par :
 - Un accompli.
 - 'كان' et indiquera, dans ce cas, la poursuite de l'effet durant une certaine période dans le passé.
كان يمشي
 - 'لم' et indiquera que l'action ne s'est pas déroulée dans le passé.
لم يرضى

III.2.Problématique de la traduction automatique des temps des verbes anglais vers l'arabe :

La traduction automatique des temps des verbes anglais vers l'arabe

pose plusieurs problèmes du fait qu'il est impossible de considérer toutes les correspondances dans un dictionnaire et la difficulté d'une représentation morpho-syntaxique. Aussi, parfois la sémantique des temps des verbes reste ambigu.

- **Exemple 1 : Present perfect continuous- Present perfect :**

Considérons l'exemple suivant :

Last year, his car **was** blue.

Ici le temps passé de l'auxiliaire 'to be ' peut correspondre à la couleur 'blue'. Dans ce cas-ci, il est compris que la voiture a été repeinte. Ou bien l'auxiliaire passé s'applique au syntagme 'his car', d'où l'on comprend que la personne a acheté une autre voiture. Ces deux interprétations relatives au temps passé de l'auxiliaire 'to be', et qui engendrent deux traductions différentes, ne peuvent être réalisées par l'ordinateur :

- كان لون سيارته أزرق.

- كانت لديه سيارة زرقاء.

C'est ce qu'on appelle 'l'indétermination sémantique du temps du verbe'. La difficulté, ici, pour l'ordinateur, réside dans le fait qu'il lui est impossible d'établir un lien entre le temps du verbe et la partie du discours qui lui correspond.

De même pour :

- Last year, I **was** ill. (1)
- Last year, I **have been** ill. (2)

Dans (1), 'was' implique que la personne était malade durant toute l'année, par contre dans (2), 'have been' implique qu'elle a été malade durant juste une certaine période de l'année.

La non-correspondance des temps entre l'anglais et l'arabe s'appréhende par le fait que le temps dans la langue anglaise s'articule et s'organise autour du présent, c'est-à-dire au moment où l'on parle. A l'inverse, le temps en langue arabe est 'atemporel', c'est-à-dire dénué de tout repère. Il s'articule autour de l'écoulement ou non de l'action. Cette différence dans l'appréciation des deux temps rend la tâche de la T.A impuissante car la logique n'est plus la même. La preuve que la langue arabe donne plus d'importance au verbe (l'action) au détriment du temps est l'existence de verbes dits 'djaAmida' (n'ayant qu'un seul temps ou aspect) et des verbes 'Naaquissa' (qui manque d'un temps). Si l'outil de mesure du 'temps' de l'action n'est plus le même, comment peut-on concevoir une traduction automatique à cela.

Exemple 1 : (present continuous, present perfect)

- The sun **has been shining** all day today.

L'usage du 'present perfect continuous' indique que l'action a débuté dans le passé et continue au moment où l'on parle. La durée de l'action s'étend, donc, du passé jusqu'au moment où l'on parle et risque de continuer. Mais existe-t-il en langue arabe un temps sémantiquement correspondant ? l'on ne pourrait traduire que par :

- ما زالت الشمس تسطع.

Qui correspond à :

- The sun **is** still **shining**. (present continuous)

Ce qui veut dire qu'on était obligé de rajouter un autre mot 'مازال' au verbe arabe 'يسطع' et malgré cela la traduction reste approximative.

De meme:

- Today he **has been playing** tennis for hours.

Le temps utilisé, ici, indique que l'action vient juste de se terminer. La langue arabe dispose-t-elle d'un temps équivalent ou d'une construction verbale similaire ? le traducteur humain, et non pas la machine, va se heurter à des difficultés de précision temporelle :

- لقد لعب التنس لساعات نهار اليوم.

Le verbe 'لعب' ne correspond nullement à la valeur temporelle de 'has been playing' car il ne permet pas au lecteur de comprendre la réalité interne de l'action.

Si parfois l'adjonction de particules, de verbes ou autres compléments de temps, au verbe arabe permet de résoudre le problème d'équivalence avec le temps dans la langue anglaise, d'autres fois cela reste insuffisant :

- It **has been raining**.

Cette phrase suggère que le lieu est encore mouillé. Si elle est traduite par :

- تساقطت الأمطار أو كان المطر يتساقط

Le lecteur arabe sera traversé par l'idée qu'il y a assez longtemps qu'il a

plu et que le lieu est sec. Ce qui veut dire, qu'en langue arabe, que la durée entre le moment du déroulement de l'action (la tombée de la pluie) et le moment où l'on parle est ambigu et pas du tout clair. Mais voyons ce qui se passe si on rajoute un autre verbe, absent dans la phrase anglaise :

- توقفت الأمطار منذ لحظات عن التساقط.

La traduction devient du coup correcte.

Cette différence dans la sémantique des temps des verbes entre l'anglais et l'arabe et la difficulté de faire une correspondance automatique réside dans la difficulté de répertorier tous les cas de figure.

- My brother **has drunk** his cup of tea (1)
- My brother **has been drinking** his cup of tea. (2)

Dans la première phrase le temps du verbe (present perfect) indique que l'action s'est réalisée complètement (il ne reste rien dans le verre), par contre, dans la deuxième phrase, le temps du verbe (present perfect continuous) indique que l'action ne s'est pas complètement réalisée (il reste du café dans le verre). Comment la langue arabe peut-elle prendre en charge le fait que l'action est complètement réalisée ou non ? si l'on traduit par :

- شرب أخي فنجانته من الشاي.

'شرب' n'a pas la même charge sémantique que 'has drunk' et 'has been

drinking’.

Le temps ‘present perfect continuous’ est parfois utilisé dans l’emphase lorsqu’il est associé à un complément du temps :

- I **have been drinking** tea **all the morning**.

La phrase est, dans ce cas, non uniforme sémantiquement, alors que la traduction arabe l’est généralement, c’est-à-dire que chaque mot arabe est sémantiquement réflexif :

- شربت الشاي طوال هذا الصباح.

On sent que la traduction est insuffisante parce qu’il n’y a rien qui met en relief la pertinence de la durée de l’action dans la phrase anglaise. Dans ce sens, la traduction arabe reste ambiguë, l’on ne comprend pas quelle partie de la phrase est importante : ce qu’on boit (le thé) ou la durée de l’action (toute la matinée).

De même pour :

- I **have drunk five cups of tea** this morning.

le temps ‘present perfect’ attire notre attention sur le nombre de verre de thé bu. L’emphase est, donc, sur le nombre alors que l’action et sa durée seront secondaires. Si l’on traduit par :

- شربت خمس فناجين من القهوة هذا الصباح.

On constate que le temps du verbe arabe ne renvoie pas à l’importance du nombre de verre bu.

Ainsi ces deux temps de la langue anglaise ‘present perfect’ et ‘present

perfect continuous' peuvent avoir un renvoi sémantique qu'on retrouve pas dans le verbe arabe.

- **Exemple 2 : 'Future tense' :**

Le futur dans la langue anglaise est exprimé essentiellement par 'to be+ going to' ou 'shall/will'. La sémantique de chacun d'eux est différente et propre : 'going to' indique le futur proche ou l'imminence de l'action dans un futur proche :

- **It's going to** rain.

Qui veut dire que la pluie va forcément tomber (du à la presence de nuages par exemple). La traduction en langue arabe nécessite l'adjonction de 'س' au verbe 'يسقط' :

- سيسقط المطر.

Malgré cela, le sens demeure imprécis et vague : serait-ce maintenant ou dans une heure par exemple ? le traducteur sera donc obligé de rajouter par exemple 'بعد قليل' qui reste insuffisante et qui ne suggère pas l'imminence de l'action. La traduction la plus proche sera alors :

- سيسقط المطر حتما بعد قليل.

'going to' a nécessité de rajouter en plus de la particule 'س', un adverbe 'حتما' et un complément du temps 'بعد قليل'.

L'usage de 'shall/will' exprime la volonté ou le désir d'exécuter une action dans le futur.

- **I will help** you.

Si la phrase est traduite par 'سأساعدك', le sens reste confiné dans

l'intention et non dans la promesse et la volonté de le faire.

III.3.Conclusion :

A travers les exemples qu'on vient de voir (on peut en citer d'autres), il s'avère que le temps du verbe anglais peut avoir plusieurs significations : il indique le moment de la réalisation de l'action, mesure la durée de sa réalisation, la façon de sa réalisation et affecte d'autres unités de la phrase. Par contre, on remarque que le verbe arabe reste 'déficient' sémantiquement. Le traducteur est obligé alors de lui joindre des particules, des compléments, des adverbes...

En général, la différence entre les deux est vaste : le verbe anglais est descriptif (action+moment de l'action+durée de l'action), alors que le verbe arabe est informatif (juste l'action).

Cet état de fait engendre plusieurs conséquences qui peuvent être difficiles à surmonter lors du processus de traduction automatique :

- L'indétermination sémantique.
- L'indétermination temporelle.

La machine ne peut agir et traiter que ce qui est explicite formellement, c'est-à-dire ce qui est écrit ou graphique, par contre l'aspect implicite (indétermination temporelle et sémantique) nécessite un formalisme rigoureux allant de la conception de dictionnaires spéciaux, à la confection d'algorithmes pouvant lire et désambiguer tous les cas de figure.

BIBLIOGRAPHIE

ABEILLE, A- (1993) : *Les nouvelles syntaxes. Grammaires a" unification et analyse du français*, Paris, Armand Colin.

ABU AL-CHAY, N. (1988) : *Un système expert pour l'analyse et la production des verbes arabes dans une perspective d'enseignement assisté par ordinateur*, Thèse en sciences de l'information, Université Lyon-1.

AMMAR, S. et J. DICHY : *Al-ka:milfi: t-tasri:f, 12.000 verbes arabes, formes et emploi*.

COHEN, D. (1961 /70) : «Essai d'une analyse automatique de l'arabe», 1961, *TA information*, in D. Cohen, *Études de linguistique sémitique et arabe*, Paris, Mouton, 1970.

DESCLES, J.-P. et al. (1983) : *Conception d'un synthétiseur et d'un analyseur morphologique de l'arabe, en vue d'une utilisation en enseignement assisté par ordinateur*, Rapport rédigé à la demande du Ministère français des affaires étrangères (sous-direction de la politique linguistique).

DICHY, J. (1987) : «The SAMIA Research Program, Year Four, Progress and Prospects», *Processing Arabie Report*, n° 2, T.C.M.O., Université catholique de Nimègue, pp. 1 -26.

DICHY, J. (1990) : *L'écriture dans la représentation de la langue : la lettre et le mot en arabe*. Thèse d'État en linguistique, Université Lumière Lyon-2.

DICHY, J. (1993): «Knowledge-system Simulation and the Computer-aided Learning of Arabie Verb-form Synthesis and Analysis», *Processing Arabie Report*, 6/7, T.C.M.O., Université Catholique de Nimègue, pp. 67-84, 92-95.

DICHY, J. (1995) : «Simulation informatique du langage naturel et compatibilité connaissances-processus : la modélisation des connaissances du niveau du mot en arabe», *Journée scientifique sur la génération et l'analyse morphologique de l'arabe*, Rabat, le 20

juin 1995, Institut d'études et de recherches pour l'arabisation.(à paraître)

DICHY, J. et M. O. HASSOUN (dir.) (1989) : *Simulation de modèles linguistiques et enseignement assisté par ordinateur de l'arabe — travaux SAMIA 1*, Paris, Conseil international de la langue française.

DITTERS, E. (1992) : *A Formal Approach to Arabic Syntax: The Noun Phrase and the Verb Phrase*, PhD, Catholic University of Nijmegen.

FRADIN, B. (1994): «L'approche à deux niveaux en morphologie computationnelle et les développements récents de la morphologie», B. Fradin (dir.), *La morphologie computationnelle, TAL (Traitement automatique des langues)*, 34 (2), Paris, Association pour le traitement automatique des langues, pp. 9-48.

HASSOUN, M. O. (1987) : *Conception d'un dictionnaire pour le traitement automatique de l'arabe dans différents contextes d'application*. Thèse d'État en sciences de l'information, Université Lyon-1.

LELUBRE, X. (1993) : «Courseware for the Theory and Practice of Arabic Conjugation», *Processing Arabic Report*, 611. T.C.M.O., Université catholique de Nimègue, pp. 85-89, 92-95.

LE NY, J.-F. (1989) : *Science cognitive et compréhension du langage*, Paris, PUE

LYONS, J. (1978/90) : *Sémantique linguistique*, 1^{re} édition 1978, traduction française de J. Durand et D. Boulonnais, Paris, Larousse, 1990.

McCARTHY, J. J. and A. S. PRINCE (1990) : «Foot and Word in Prosodic Morphology: the Arabic Broken Plural», *Natural Language and Linguistic Theory*, 8 (2), pp. 209-283.

MEL'CUK, I. A. (1982a) : *Towards a Language of Linguistics, A System of Formal Notions for Theoretical Morphology*, München, Wilhem Fink.

MEL'CUK, I. A. (1982b) : «Élaboration d'un langage formel pour la morphologie», C. Bertaux, J.-P. Desclés, D. Dubarle *et al.*. *Linguistique et mathématiques. Peut-on construire un discours cohérent en linguistique ?*, Berne, Peter Lang, pp. 99-119.

ROMAN, A. (1990) : *Grammaire de l'arabe*, Paris, PUF, coll. «Que sais-je ?».

SABAH, G. (1988/89) : *L'intelligence artificielle et le langage*, vol. 1 : *Représentation des connaissances* (1988), vol. 2 : *Processus de compréhension* (1989), Paris, Hermès.

CONCLUSION GENERALE

De l'avis de beaucoup des spécialistes les systèmes d'avenir combineront des méthodes de règles assez traditionnelles et des méthodes statistiques et basées sur les exemples. Ils seront hybrides. Mais de quelle façon? Dans une optique possible, on peut envisager que les méthodes linguistiques des systèmes indirects fourniront la base pour l'application des méthodes basées sur les banques de connaissances, sur les données statistiques et sur les exemples de textes traduits. La base linguistique fournira les analyses syntaxiques et sémantiques assez simples, l'essentiel du transfert et l'information syntactique pour la génération. A la base on verra probablement les formalismes d'unification et les grammaires basées sur contraintes. Les autres méthodes fourniront une désambiguïsation plus souple et les informations nécessaires pour le transfert lexical et pour la production des textes idiomatiques.

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

- les règles seront moins ambitieux que celles des systèmes indirects de transfert et d'interlingue;
- l'analyse syntaxique sera limitée à la reconnaissance des structures superficielles, des composantes de phrase et des relations de dépendance;
- il n'y aura presque aucune analyse profonde des relations logiques;
- l'analyse sémantique sera limitée à l'identification des rôles d'éléments dans les phrases: l'agent, l'instrument, etc.
- les informations lexicales seront extraites principalement des sources régulières comme des dictionnaires généraux destinés au grand public; par conséquent elles n'indiqueront que des catégories syntactiques et peut-être des traits sémantiques peu raffinés.
- ces traits sémantiques assez rudimentaires ne seront utilisés que pour une désambiguïsation initiale.

- les règles de transfert lexical et structural s'appliqueront peut-être à des représentations peu profondes (quoique moins simples que dans le système d'IBM).

Quant aux méthodes encore assez nouvelles:

- des exemples de traduction, stockés dans une banque de textes bilingues alignés, seront utilisés pour aider la désambiguïsation plus délicate dans l'analyse de la langue source et pour choisir les équivalents dans la langue cible

- des données statistiques sur les collocations lexicales et les fréquences du vocabulaire monolingues aideront l'analyse syntaxique et sémantique des phrases, la désambiguïsation monolingue, et le choix des phrases idiomatiques dans la langue cible.

- des données sur les probabilités des équivalences bilingues seront utilisées pendant le transfert lexical.

- des banques de connaissances des domaines en question aideront la désambiguïsation monolingue et interlingue.

On peut anticiper aussi d'autres développements importants:

- l'emploi des méthodes de feedback pour améliorer les grammaires (ou le fond de règles) et les lexiques monolingues et bilingues

- les recherches plus approfondies sur les problèmes de discours et de style

- l'intégration de la traduction automatique dans les systèmes pour la production, la transmission et le management de documents dans les bureaux (à l'exemple des postes de travail pour traducteurs)

L'Utilisation des systèmes

Aujourd'hui, les utilisateurs et les chercheurs sont tous réalistes. Le système de traduction complètement automatique qui produira les textes idiomatiques

comparables aux traductions humaines n'est qu'un rêve pour ceux qui n'en ont pas l'expérience. On ne croit pas non plus à la réalisation pratique et économique de systèmes à l'usage général, qui traitent une grande gamme de domaines. La recherche est centrée sur le développement de systèmes limités à un sous-langage ou à un domaine technique spécial. Dans les conditions propices, les systèmes loins d'être parfaits peuvent être utilisés avec profit et succès. Sans doute tout le monde aspire à des systèmes d'une meilleure qualité, mais on ne les attend pas dans un proche avenir. Il faut rappeler que c'est toujours les systèmes expérimentaux qui expérimenteront avec des techniques nouvelles. On n'attend pas l'arrivée d'aucun système commercial basé sur ces méthodes de la "troisième génération". Dans cette nouvelle époque les systèmes pour le marché seront basés sur les méthodes sûres, bien établies et bien éprouvées qui, toutefois, ne seront pas les méthodes les plus innovatrices.

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

En somme, on prévoit une période d'expérimentation variée et assez incohérente quant à la théorie. En même temps on prévoit une expansion d'utilisation des

systèmes de grande taille et celle du nombre d'utilisateurs de systèmes basés sur les ordinateurs personnels. Quelle sortes de systèmes satisferont ces nouveaux besoins? Les chercheurs devront s'adresser à ces défis, et aussi, avec d'autant plus d'urgence, à l'établissement de normes pour permettre d'évaluer et de comparer la performance, la qualité et l'efficacité des systèmes commerciaux. On attend les changements rapides dans le domaine de traduction automatique dans le proche avenir avec beaucoup d'intérêt et d'espoir.

Résumé

La Traduction automatique est sans doute l'une des applications les plus complexes du domaine du traitement automatique des langues naturelles (TALN), et ce pour deux raisons : si l'on admet l'importance des connaissances linguistiques, elle nécessite, au moins, l'analyse du texte à traduire, indispensable à l'extraction du sens, et la production d'un texte porteur du même sens en langue cible. Chaque étape est semée d'embûches.

Ainsi au niveau de l'analyse, il faut répertorier et disséquer toutes les difficultés lexicales, morphologiques et syntaxiques. Chose qui n'est pas évidente car chaque système linguistique possède ses spécificités propres. Ensuite, il faut passer à la phase de transfert où il s'agit de passer la somme d'informations (ou connaissances) récoltée d'une représentation donnée vers une autre représentation par le biais de programmes spéciaux. Enfin, il faut générer cette somme d'informations dans la langue cible.

L'objectif recherché, à travers ce mémoire, est de présenter les domaines de connaissances de la traduction automatique, les différentes étapes mises en œuvre par de tels systèmes (analyse, transfert et génération) ainsi que l'état de l'art de la T.A.

Dans ce sens, on passe en revue les différentes difficultés auxquelles est confrontée la T.A : les difficultés lexicales, syntaxiques, de génération, d'analyse et de transfert.

Le premier chapitre dresse les étapes principales dans l'histoire de la T.A. On s'attèlera sur l'évolution des principes directeurs ayant participé à forger l'institution de la T.A comme discipline scientifique inévitable. La situation, ainsi que la complexité y sont brièvement évoqués.

Le second chapitre se veut une introduction à la T.A. On présente, notamment, les généralités sur la T.A : les niveaux de traitement linguistique et les connaissances nécessaires à la T.A. On énumérera les différents systèmes actuels et la tendance de chacun d'eux.

Dans le troisième chapitre, on présente le bilan de quelques systèmes de traduction commercialisés ou réalisés, en décrivant, de façon exhaustive, les différentes phases utilisées. On se focalisera, surtout, sur les systèmes prenant en charge ou utilisant la langue arabe (le système TORJOMAN, le système Aleth Trad...).

Dans le quatrième chapitre est présenté quelques difficultés inhérentes à la T.A telles que la délimitation des unités de traduction sachant que le mot n'est pas toujours une unité valable, la conception d'un dictionnaire interne....

Le cinquième chapitre est consacré à la problématique de la T.A et la langue arabe. On y évoque quelques problèmes posés, surtout ceux ayant trait à la morphologie et la morpho-syntaxe sachant que la langue arabe est une langue à forte morphologie. On évoquera, en particulier, la structure et la tentative de la représentation du mot arabe, ainsi que la structure verbale arabe.

Enfin, une conclusion est inférée sous forme de résumé sur l'espoir futur qu'inspire la T.A à travers des perspectives d'avenir.

Mots clés :

Traduction automatique; Linguistique; Unité de traduction; Dictionnaire interne; Système de traduction.