

enssib
Ecole Nationale Supérieure
des Sciences de l'Information et des Bibliothèques

D E A

Sciences de l'Information et de la Communication

OPTION :
Systèmes d'Information Documentaire

Note de Synthèse de DEA

Traduction Assistée par Ordinateur
du français vers l'arabe :
état de l'art

Auteur :
Sahbi SIDHOM

Sous la Direction :
Mohamed HASSOUN
&
Joseph DICHY

1996

Université Lumière
Lyon 2

Ecole Nationale Supérieure
des Sciences de l'Information et des Bibliothèques

Université Jean Moulin
Lyon 3

SOMMAIRE

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

Partie 1 : Tour d’horizon

I. La Traduction Automatique:.....	2
I.1. Introduction:.....	2
I.2. Les connaissances nécessaires pour la TA:.....	4
I.2.1. Les dictionnaires:.....	4
I.2.2. Les connaissances syntaxiques :.....	5
I.2.3. Les connaissances sémantiques:.....	6
I.2.4. Les connaissances contextuelles:.....	8
II. Etat de l'art de la TA ou la TAO :.....	11
II.1. Première génération:(1950-1962).....	11
II.2. Deuxième génération :(1960-1970).....	11
II.2.1. Evolution de l’architecture des systèmes de deuxième génération:.....	13
II.2.1.1. Le modèle à transfert:(type <i>PIVOT</i>)......	14
II.2.1.2. Le modèle à interlangues:(type <i>SHALT2</i>)......	14
II.3. Troisième génération:.....	15
II.3.1. Architecture d’un système de troisième génération:.....	15
III. Description de quelques systèmes de TA:.....	16
III.1. Le système ALT-J/E:.....	16
III.1.1. Les dictionnaires:.....	16
III.1.2 Le schéma général du système:.....	17
III.1.3. Fondement théorique du système:.....	17
III.1.4 Mécanisme de traduction du système:.....	18
III.2 Le système ARIANE:.....	19
III.2.1 Le processus de traduction:.....	19
III.2.1.1 Analyse:.....	20
III.2.1.2 Transfert:.....	20
III.2.1.3 Génération:.....	20
III.2.2 Evolution d’ARIANE: <i>ARIANE-G5, 1988</i>	21
III.2.3 Utilisation de l’aspect multilingue d’ARIANE:.....	22
III.2.4 Application - traducteur de l’anglais vers l’arabe :.....	23
III.3. Le système AlethTrad:.....	25
III.3.1. L’analyse des textes:.....	25
III.3.2. Traduction , révision et validation:.....	25
III.3.3. Traduction multi-niveaux:.....	27
III.3.4 Architecture d’AlethTrad:.....	31
III.4 Le projet METAL:.....	32
III.4.1 Historique et évolution de METAL:.....	32
III.4.2 Le système METAL:.....	32
III.4.3 Les outils du développement lexical:.....	36

III.4.4 Evolutions:.....	37
III.5 Le système TOSHIBA :.....	38
III.5.1 Introduction:.....	38
III.5.2 Les caractéristiques du système:.....	38
III.5.3 Configuration du système:.....	38
III.5.4 La méthode de traduction :.....	39
III.5.4.1 Analyse morphologique - dictionnaires :.....	39
III.5.4.2 Analyse Syntaxique:.....	41
III.5.4.3 Analyse Sémantique:.....	43
III.5.4.4 Transfert structural:.....	44
III.5.4.5 La génération:.....	46
III.5.4.5.1 La génération syntaxique:.....	46
III.5.4.5.2 La génération morphologique:.....	46
III.6. Le système TORJOMAN:.....	49
III.6.1. Introduction:.....	49
III.6.2. Aspects fonctionnels :.....	49
III.6.3. Aspects techniques :.....	49
III.6.3.1. La compilation des dictionnaires :.....	50
III.6.3.2. L'analyse de l'anglais :.....	51
III.6.3.3. Le transfert et la génération :.....	51
III.6.4. Caractéristiques du système TORJOMAN :.....	53
III.6.5. Architecture du 'moteur' de traduction :.....	54
III.6.6. Remarques sur le système TORJOMAN :.....	55
CONCLUSION.....	40
BIBLIOGRAPHIE.....	58

INTRODUCTION

Le but de la recherche en ingénierie linguistique est d'ajouter de la valeur aux résultats déjà obtenus, de manière à consolider la structure de l'industrie télématique.

La vocation des projets de recherche en ingénierie linguistique est de perfectionner les modèles de dialogue homme-machine, ainsi que les technologies et les méthodes de **T**raitement **A**utomatique du **L**angage **N**aturel (TALN) .

Ce qui permet, par un traitement plus robuste d'améliorer la compréhension et la **T**raduction **A**utomatique (TA) des énoncés linguistiques courants, tant oraux qu'écrits.

L'informatique peut être accablée de bien des défauts, mais l'une de ses principales vertus est d'être souvent plus performante que les humains, lorsqu'un mécanisme peut être décrit par une algorithmique précise qui rend possible l'automatisation.

Depuis les années 50 de nombreux chercheurs ont travaillé à l'élaboration de systèmes de TA. Aux Etats-Unis, tout d'abord, l'idée de traduction est apparue comme une voie de développement dans l'utilisation des ordinateurs.

L'objectif de notre note de synthèse bibliographique est de présenter les domaines de connaissances de la TA, les différentes étapes mises en œuvre par de tels systèmes (analyse, transfert et génération) et l'état de l'art de la **T**raduction **A**ssistée par **O**rdinateur (T.A.O).

Enfin nous présenterons une description de quelques systèmes de TA commercialisés ou réalisés.

I. La Traduction Automatique:

I.1. 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)** [NIRENBURG 87b] .

La qualité du texte produit en génération LC est irréprochable et similaire au travail d'un traducteur professionnel [ARNOLD 87].

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.2. Les connaissances nécessaires pour la TA:

I.2.1. Les dictionnaires: [HASSOUN 87],[FARWELL 93]

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 allemand:

"Das Buch liegt auf dem Tisch." (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 :

<u>Allemand</u>	<u>Anglais</u>
auf	on
Buch	book
das	the
dem	the
liegt	is
Tisch	table

Le programme de traduction a fourni une substitution des mots allemands 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 l'allemand ?

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

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

I.2.2. Les connaissances syntaxiques : [LYTINEN 87], [SABAH 90a]

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 \Leftarrow NP + VP
NP \Leftarrow D + N
VP \Leftarrow 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 :

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

I.2.3. Les connaissances sémantiques:

[PITRAT 85],[RICHARD 95], [LYTINEN 87]

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 *)

).

I.2.4. Les connaissances contextuelles: [CARBONELL 87],[SABAH 90b]

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 [PITRAT 85] :

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 au paravant 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.

II. Etat de l'art de la TA ou la TAO :

Nous allons décrire l'état de l'art de la TA ou de la TAO (Traduction automatique Assistée par Ordinateur) en classant les systèmes par génération [RADWAN 94],[BOUALEM 93], [MELBY 87].

II.1. Première génération:(1950-1962)

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.

Inconvénients:

- L'accumulation des phénomènes linguistiques devient si importante, que l'insertion de nouveaux phénomènes dans les programmes, de plus en plus complexes, devient pratiquement impossible [CULLINFORD 87]. Le niveau de qualité atteint assez vite sa limite.
- Les programmes linguistiques sont directement écrits dans les langages de programmation de bas niveau .

II.2. Deuxième génération :(1960-1970)

Une autre approche pour améliorer la qualité de la TAO , consiste en la division du processus de traduction en trois phases logiques :

- 1- Analyse,
- 2- Transfert,

3- Génération.

Illustration par la fig.II.1 .

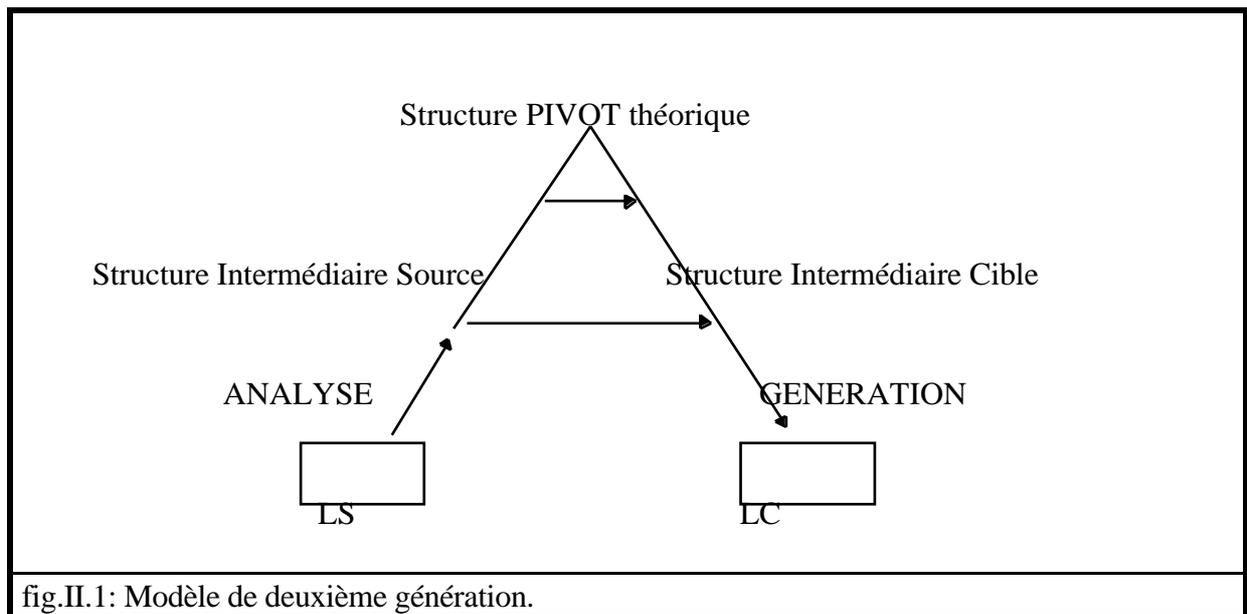


fig.II.1: Modèle de deuxième génération.

La sortie de la phase **analyse** est un descripteur structural du texte en LS.

Ce dernier étant transformé en descripteur structural équivalent en LC par la phase de **transfert** [KABBAJ 91].

Ce descripteur cible est alors transformé en texte de LC par la phase de **génération**.

Seule la deuxième phase logique fait intervenir le couple de langues ; plus l'analyse est profonde, plus la distance entre les deux descripteurs structuraux est courte [NAGAO 87]. 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 [DORR 93], [NIRENBURG 87a].

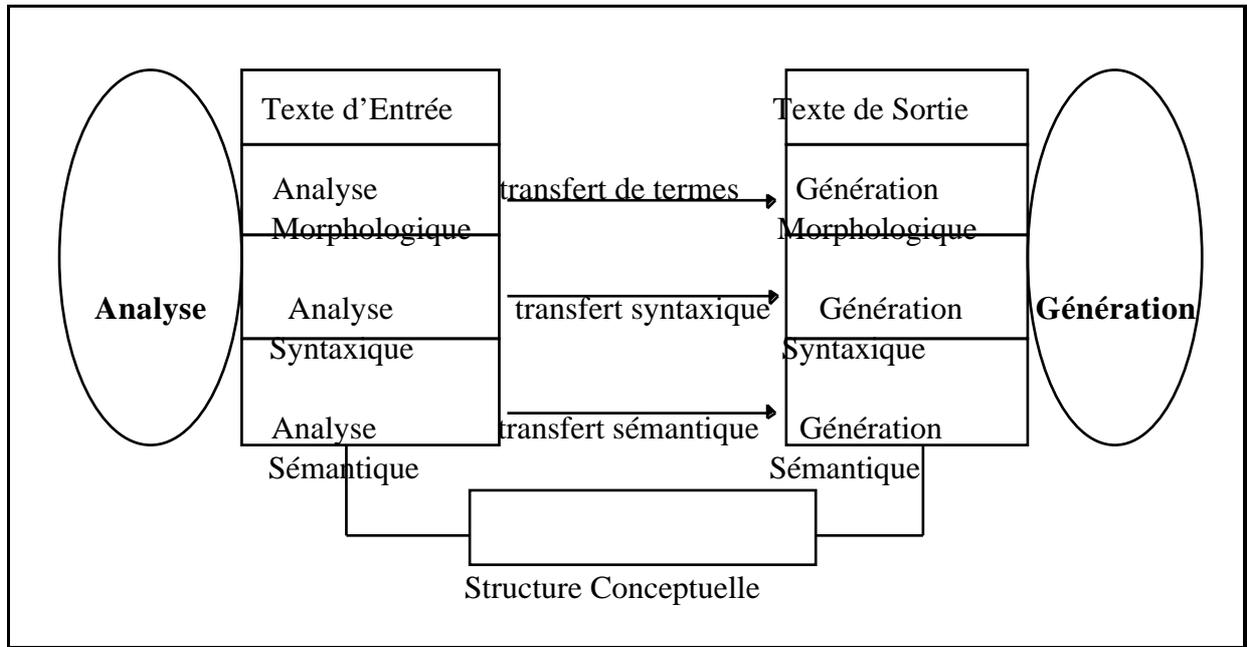
Inconvénients:

- 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.

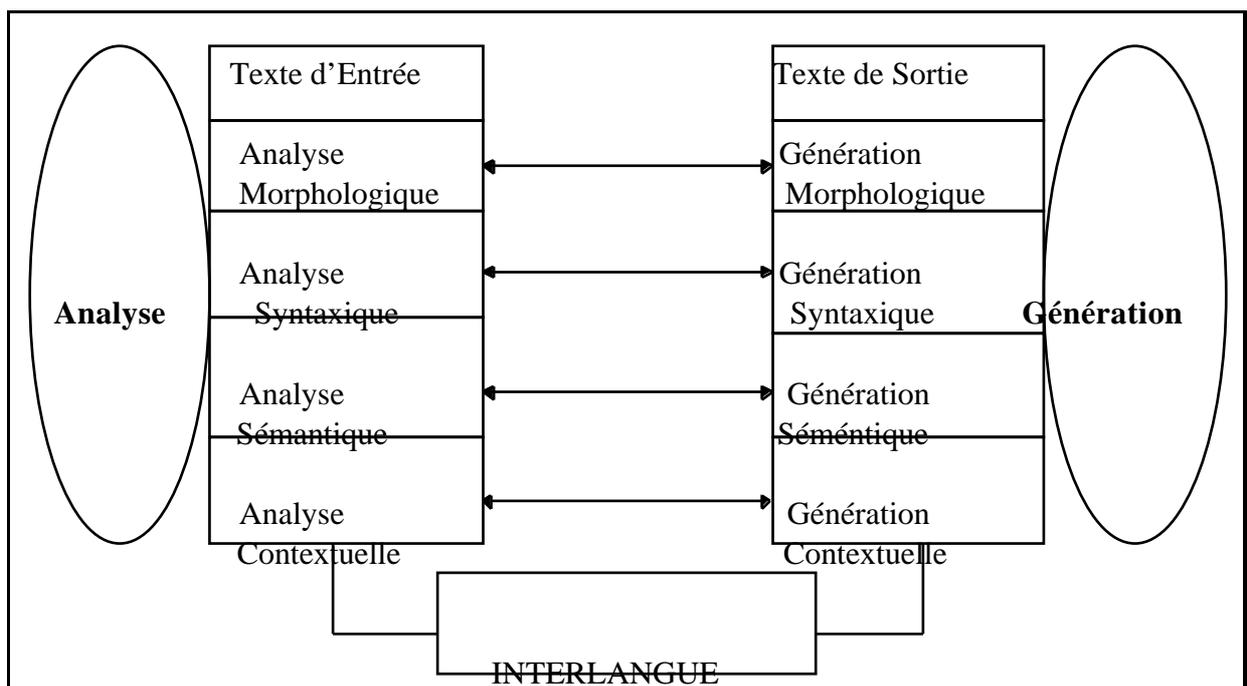
II.2.1. Evolution de l'architecture des systèmes de deuxième génération: [ABBOU 95a]

L'architecture des systèmes est comme suit:

II.2.1.1. Le modèle à transfert:(type PIVOT)



II.2.1.2. Le modèle à interlangues:(type SHALT2)



II.3. Troisième 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 [MCDONALD 87], 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 [GROSS 95].
- 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.3.1. Architecture d'un système de troisième génération:

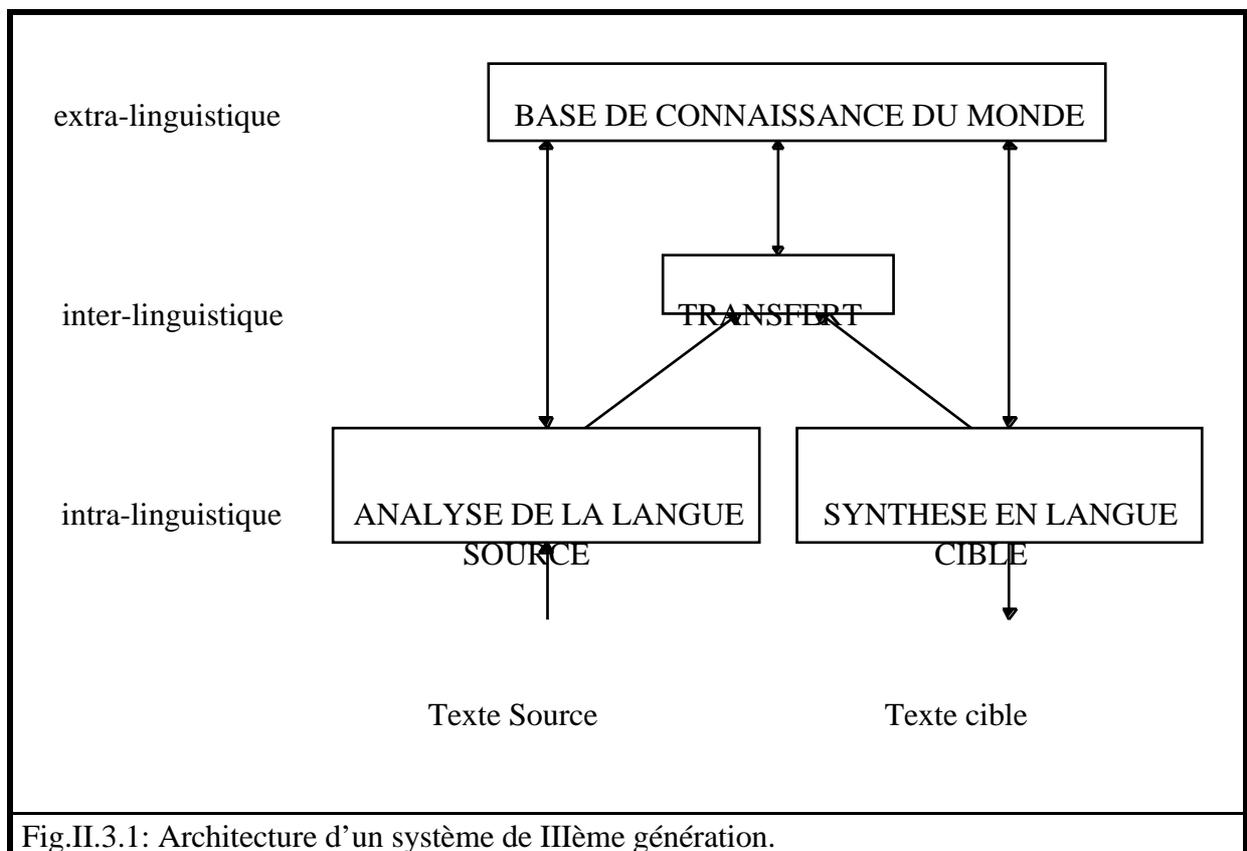


Fig.II.3.1: Architecture d'un système de IIIème génération.

III. Description de quelques systèmes de TA:

III.1. Le système ALT-J/E: [ABBOU 95b]

Développé par le NTT:Laboratoire des Sciences de la Communication au Japon. Le projet a débuté en 1985.

III.1.1. 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.

III.1.2 Le schéma général du système: (Voir Fig.III.1)

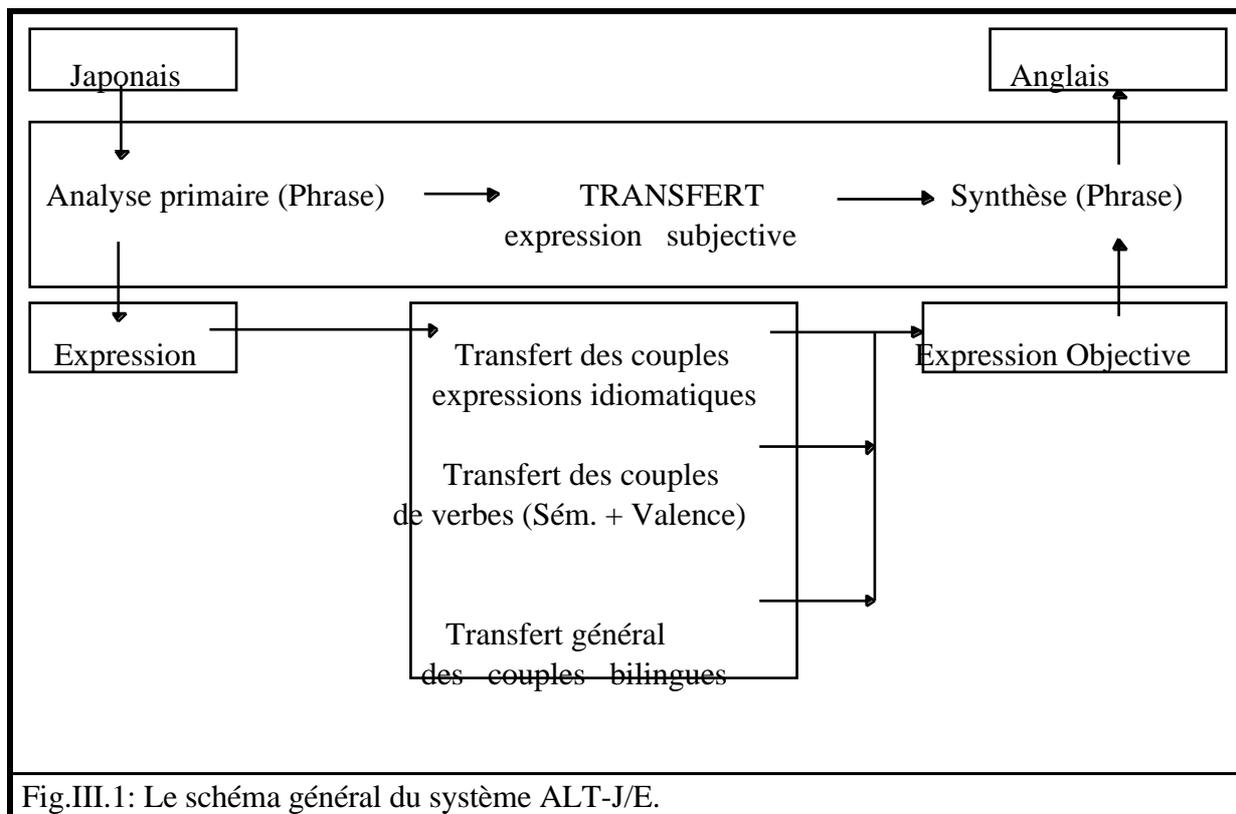


Fig.III.1: Le schéma général du système ALT-J/E.

III.1.3. Fondement théorique du système: [ABBOU 95b]

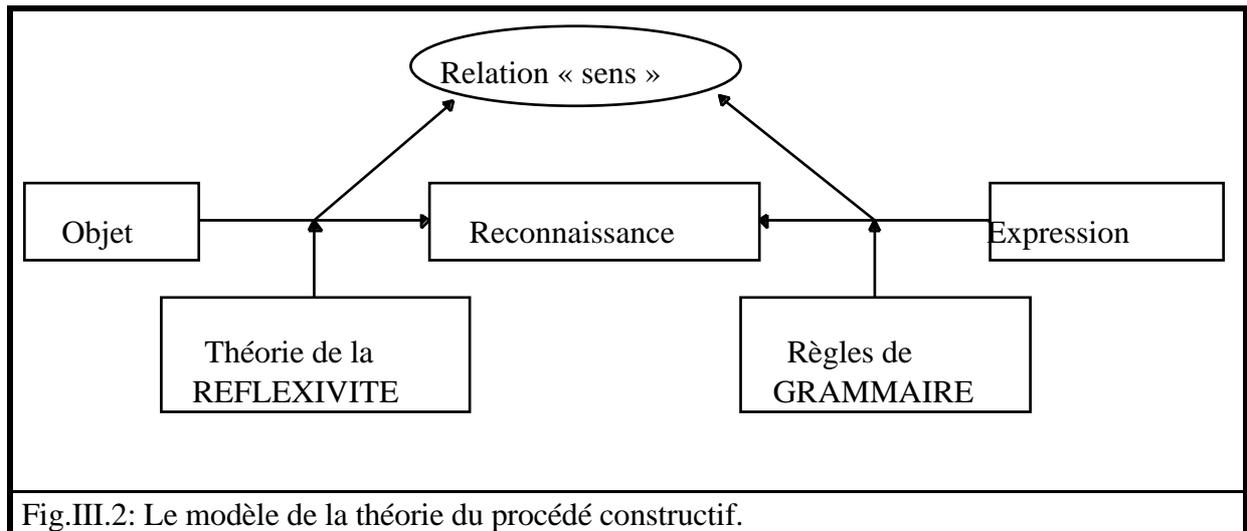
La *Théorie du Procédé Constructif* (TCP) est issue des travaux d'un grande école linguistique japonaise (cf. Pr M. Miyazaki, Japon IA N° 28).

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. (Voir fig.III.2)



III.1.4 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 oeuvre 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.

III.2 Le système ARIANE: [MONEIMNE 89],[VEILLON 70],[VAUQOIS 75]

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:

- 1- Quatre langages spécialisés pour la programmation linguistique: TRANSF, ROBRA, ATEF, SYGMOR.
- 2- 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.
- 3- Un ensemble de programmes utilitaires qui facilitent la mise au point et l'exploitation des programmes linguistiques.

III.2.1 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.		

III.2.1.1 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.

III.2.1.2 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.

b) 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.

III.2.1.3 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

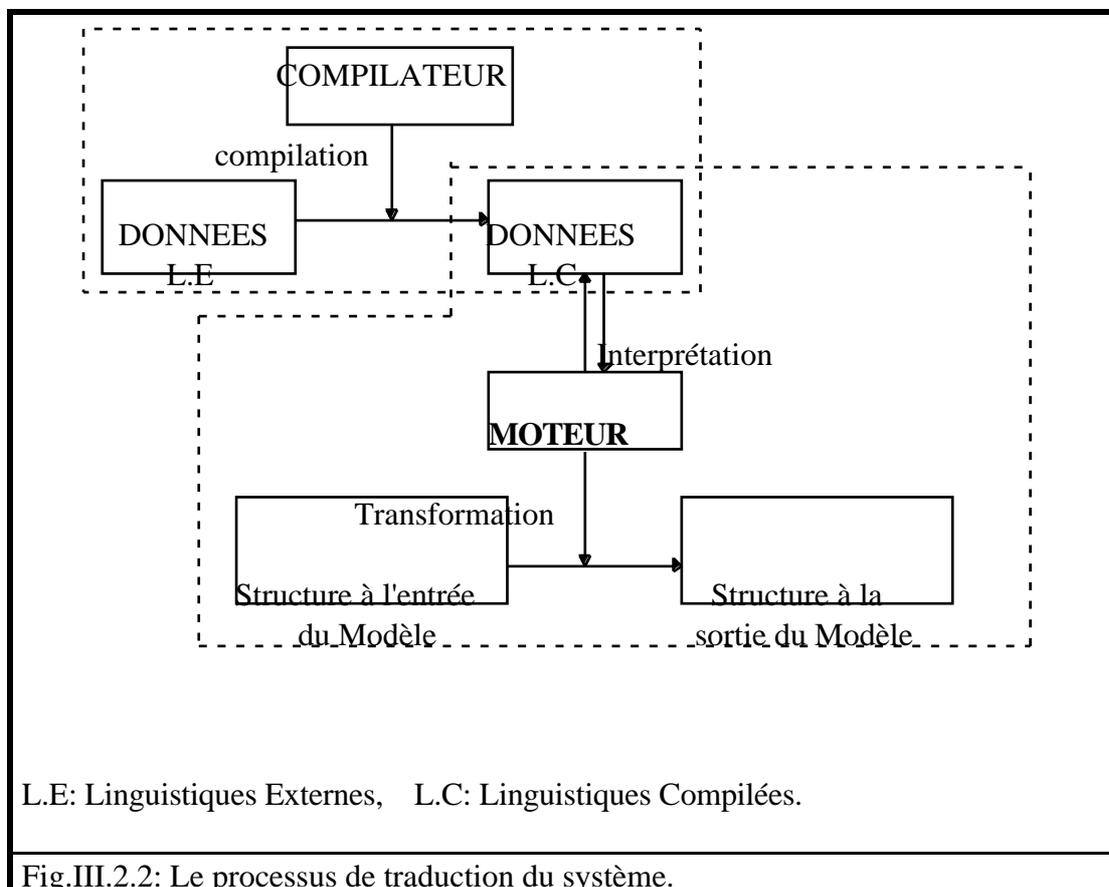
Il travaille sur les feuilles de l'arbre (ou terminaux), produit par le GS, et produit des formes morphologiques correctes constituant un texte en LC .

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.

III.2.2 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 .



III.2.3 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 .

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

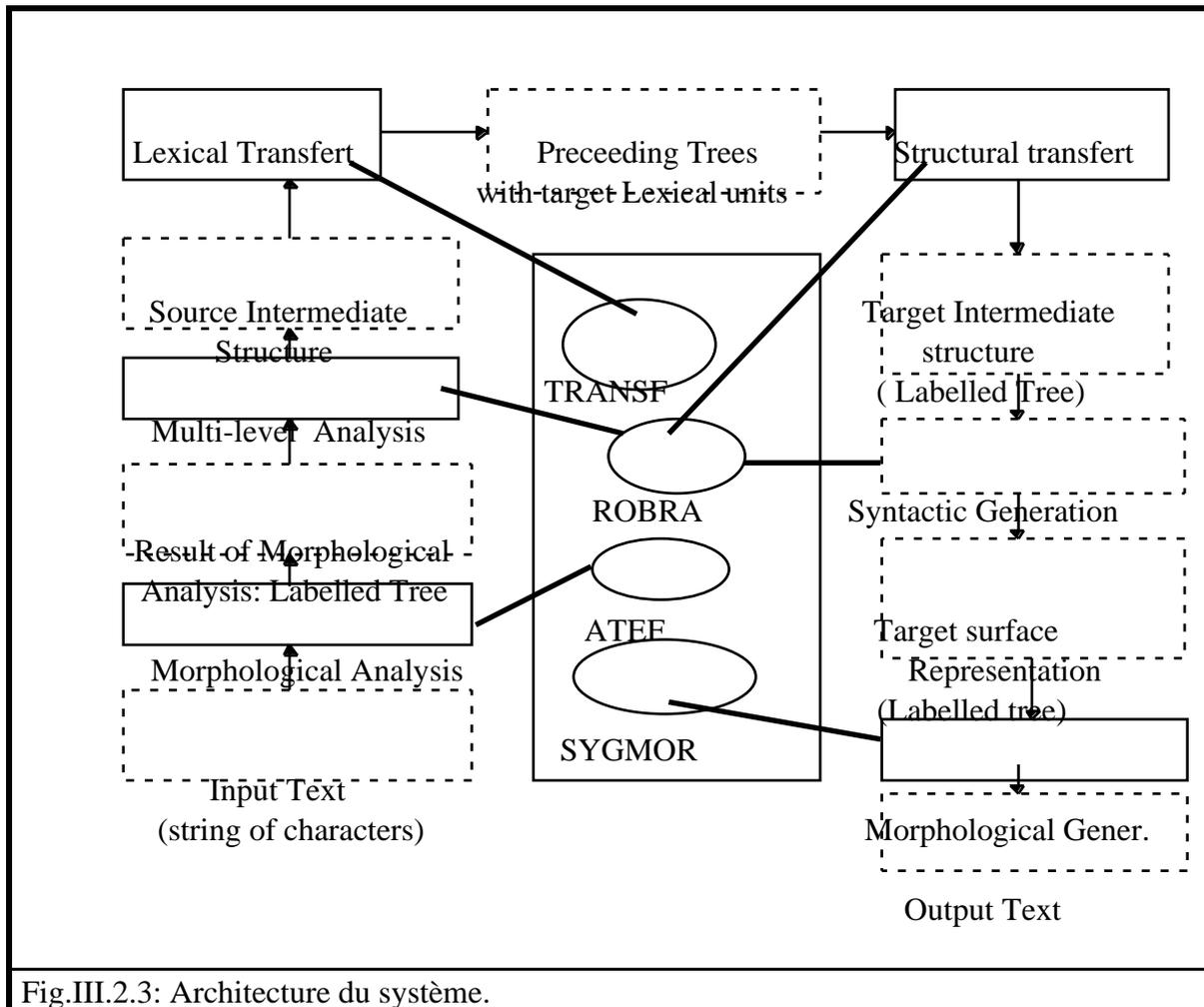


Fig.III.2.3: Architecture du système.

III.2.4 Application - traducteur de l'anglais vers l'arabe :

[MONEIMNE 89]

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.

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 traité 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. On en est bien loin !
- 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.

- les bases de traduction: dictionnaire de LS, dict. bilingues,...
- les mémoires de traduction: terminologie et segments bilingues.

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

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.

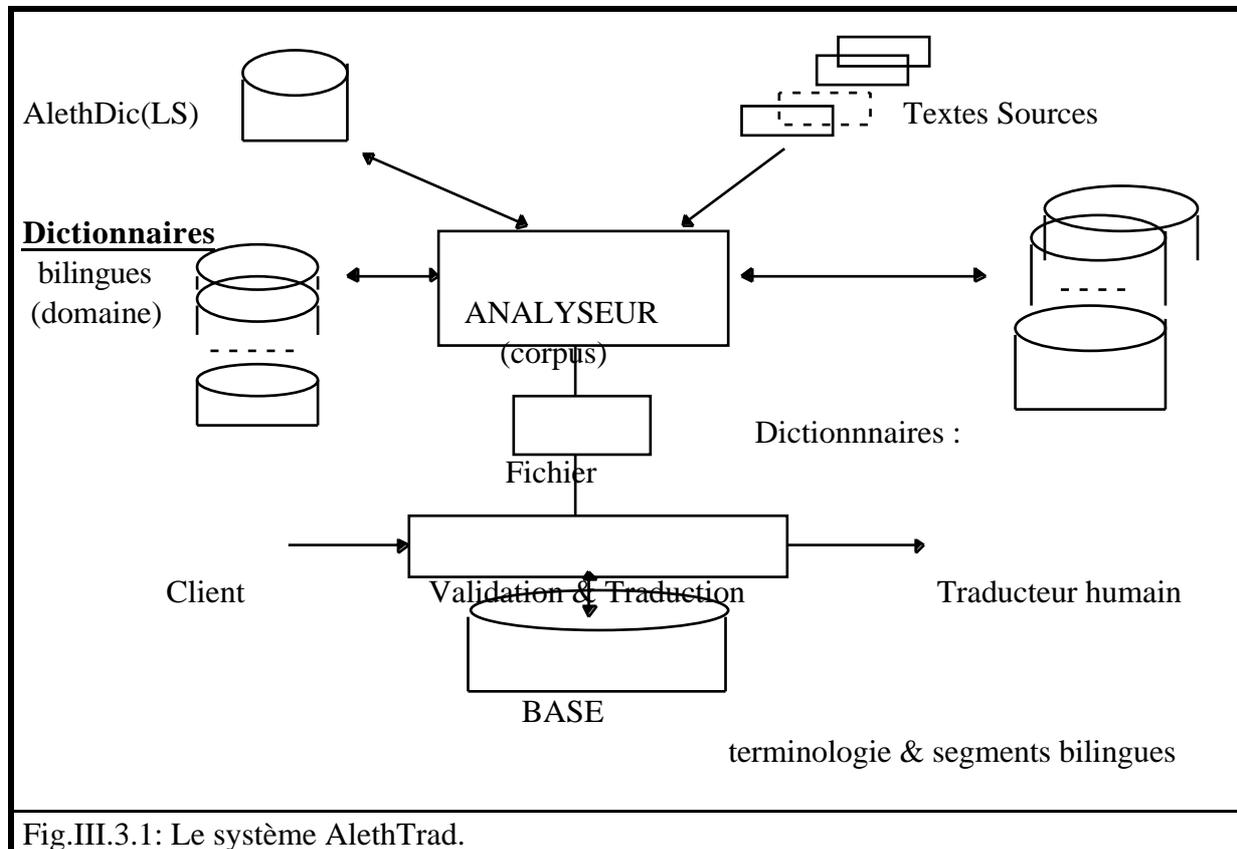


Fig.III.3.1: Le système AlethTrad.

III.3.3. Traduction multi-niveaux:

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 pharséologiques 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 .

c) 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 : niveau 2 .

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.3.4 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) .

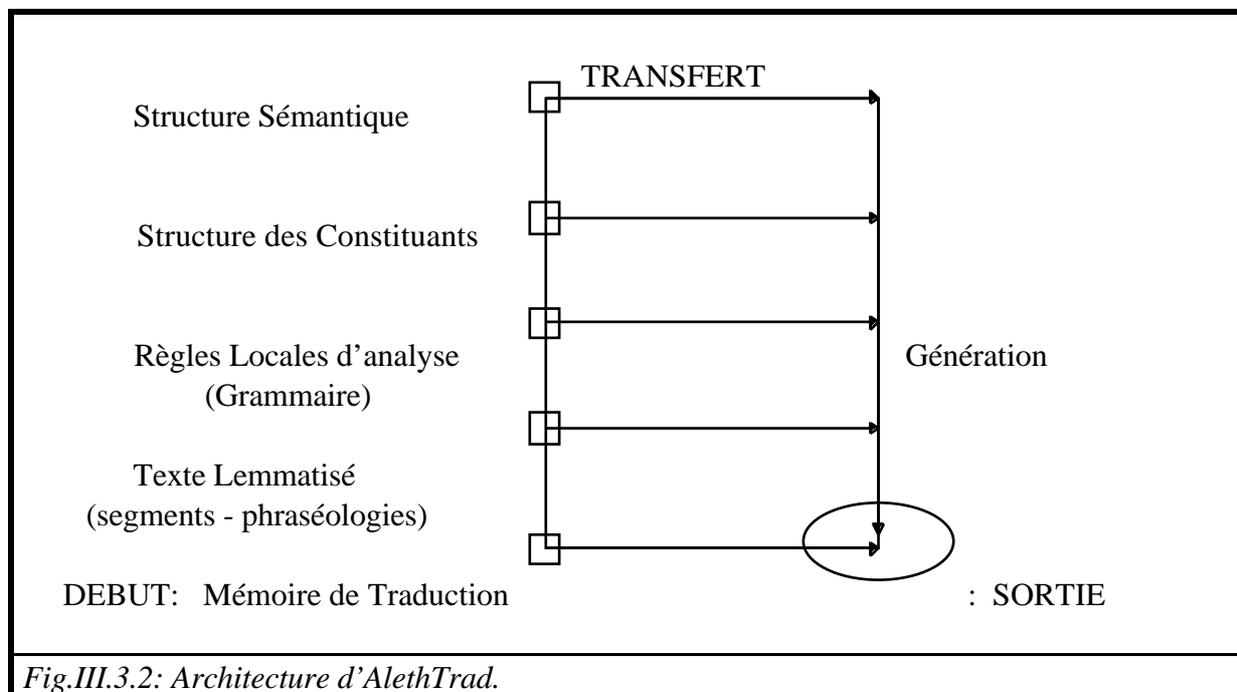


Fig.III.3.2: Architecture d'AlethTrad.

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.

III.4 Le projet METAL: [WHITE 87]

Origine: Université de Texas - USA et SIEMENS en Allemagne .

Période: 1961-1990 .

III.4.1 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

III.4.2 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 noeud 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 noeud fils au noeud 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 :

- ① : L'appel de la fonction TRANSLATE (appel à l'environ. Lisp).
- ② : La mise sous forme d'une liste de la phrase.
- ③ : Le nombre d'interprétations de l'analyseur.
- ④ : Le transfert et la génération .
- ⑤ : 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:

- ① (translate)
- ② Sentence: (Ein alter, verheirater Mann hat dieses Programm niemals geschrieben.)
- ③ 1 interpretation in 25314 ms. 286 phrases: 250 rejected.
- ④ Transfert plus generation time: 14538 ms.
- ⑤ (|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 » :

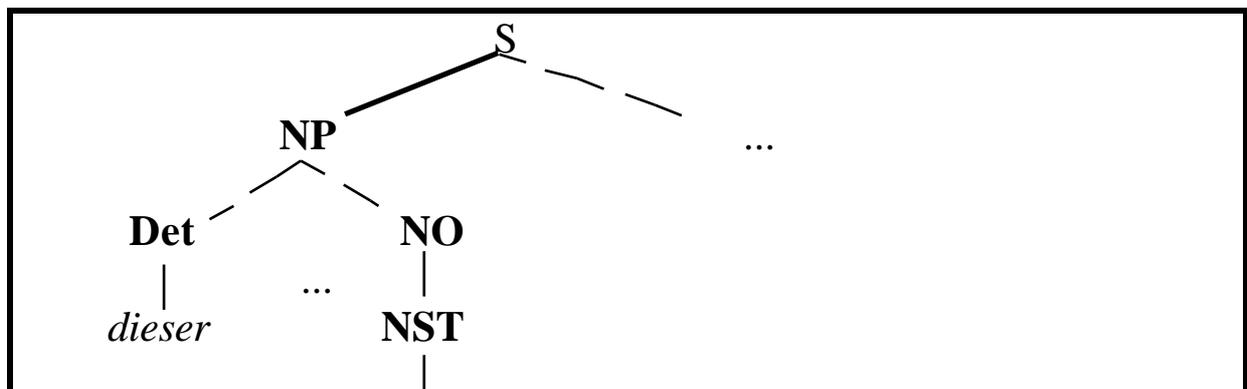
Le P-tree ou P-arbre ou arbre décoré est la forme d'interprétation étalée communément utilisée par les développeurs du système. Il apporte du sens en visualisant la phrase analysée.

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 noeuds 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](#) .



<i>Programm</i>

<i>Fig.III.4.3: Arbre d'analyse.</i>

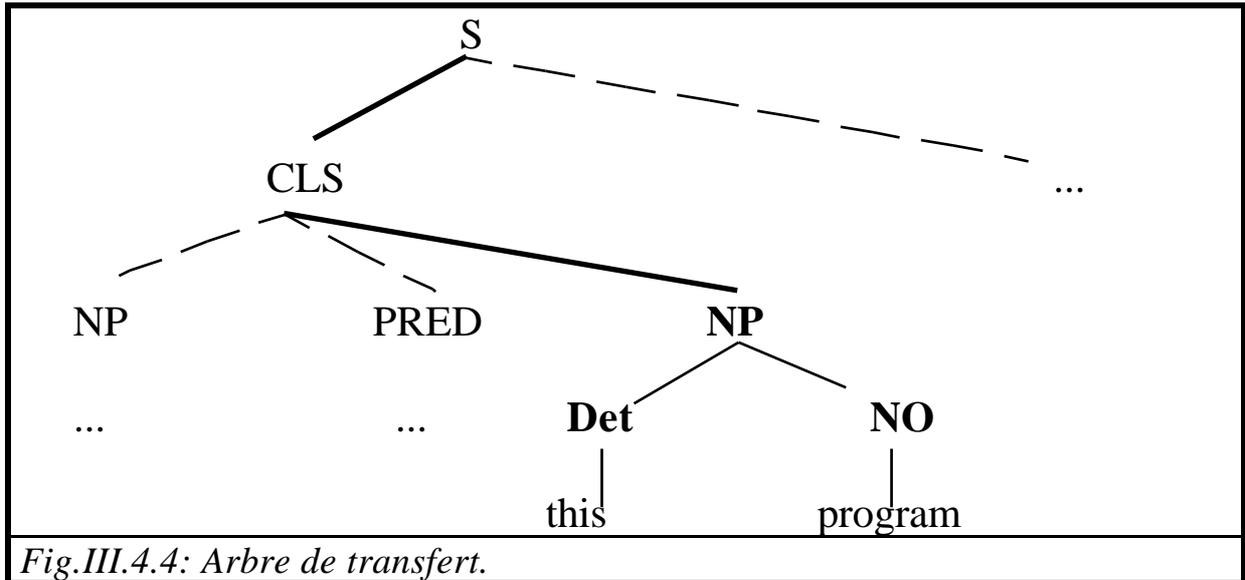


Fig.III.4.4: Arbre de transfert.

III.4.3 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.

III.4.4 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.

III.5 Le système TOSHIBA : [AMANO 89]

III.5.1 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) .

III.5.2 Les caractéristiques du système:

Le système comporte :

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

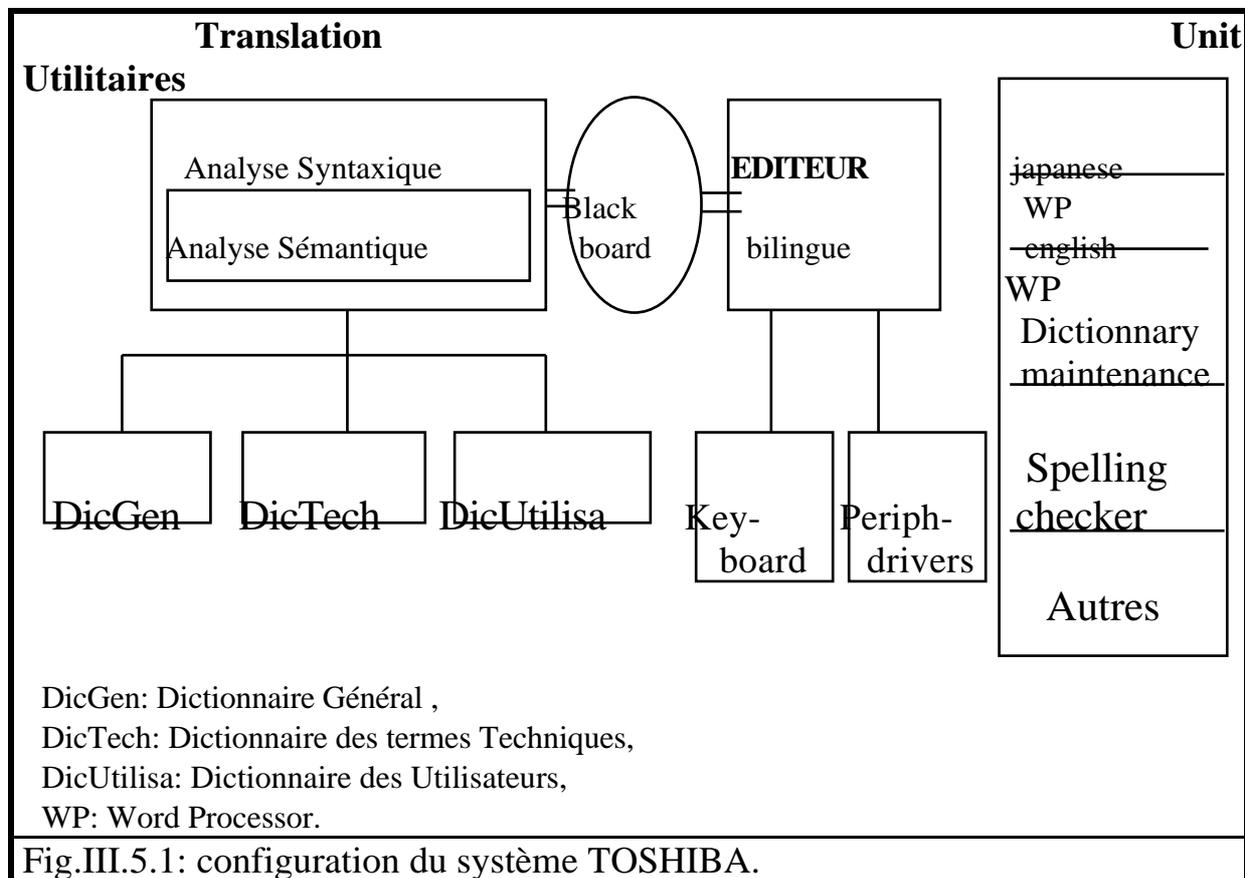
L'analyseur sémantique représente la signification de la phrase comme une structure conceptuelle de celle de la structure syntaxique et de la signification des mots qui apparaissent dans la phrase.

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 .

III.5.3 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 .



III.5.4 La méthode de traduction :

III.5.4.1 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 .

III.5.4.2 Analyse Syntaxique: ASy

Les analyses syntaxique et sémantique sont séparées et non complètement indépendantes : elles fonctionnent séquentiellement et procèdent dans un mode interactif. (Fig.III.5.3)

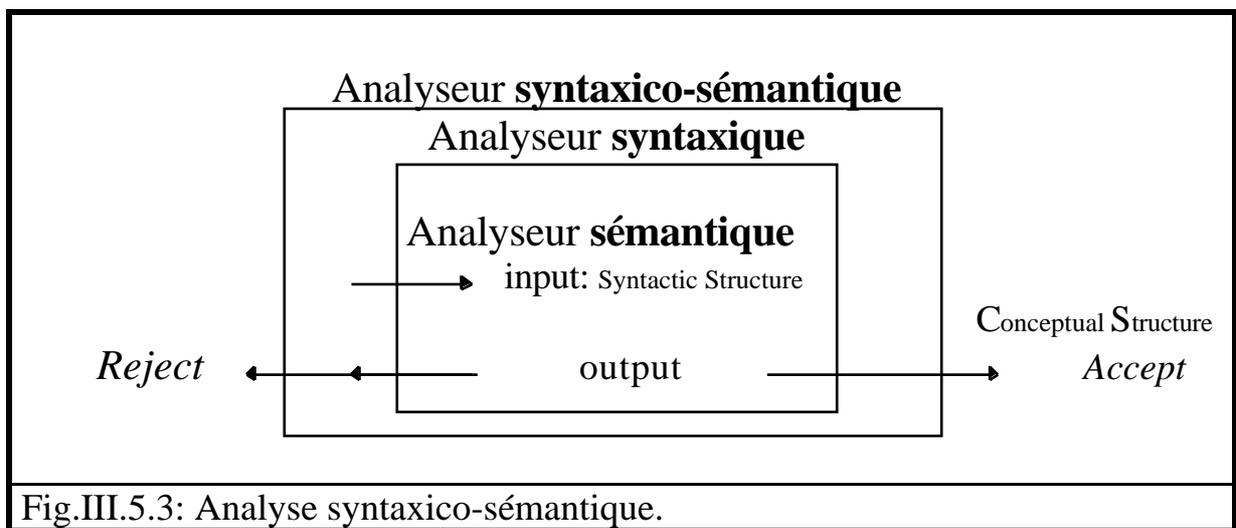


Fig.III.5.3: Analyse syntaxico-sémantique.

Les traits de l'analyse syntaxique du système sont :

- 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.

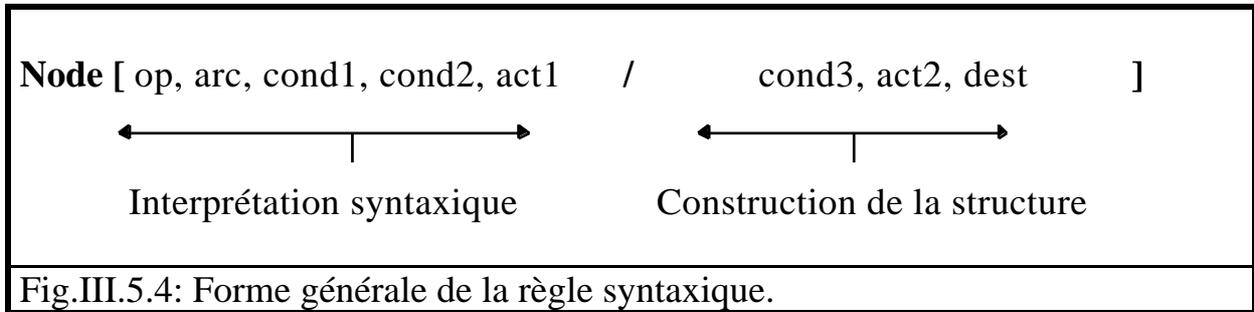
Si l'ambiguïté persiste, l'analyse sémantique construira une structure conceptuelle plausible et résoudra implicitement les ambiguïtés.
- 2- L'analyse syntaxique est purement syntaxique et les règles syntaxiques ont des conditions sémantiques .
- 3- L'analyse syntaxique est composée de deux parties :
 - 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.



III.5.4.3 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:

$$\text{MP} = \text{TP} ; \text{COND} ; \text{ACT} ; \text{CTRL} .$$

où :

MP « Matching Pattern »: correspondant à l'arbre de la structure syntaxique.

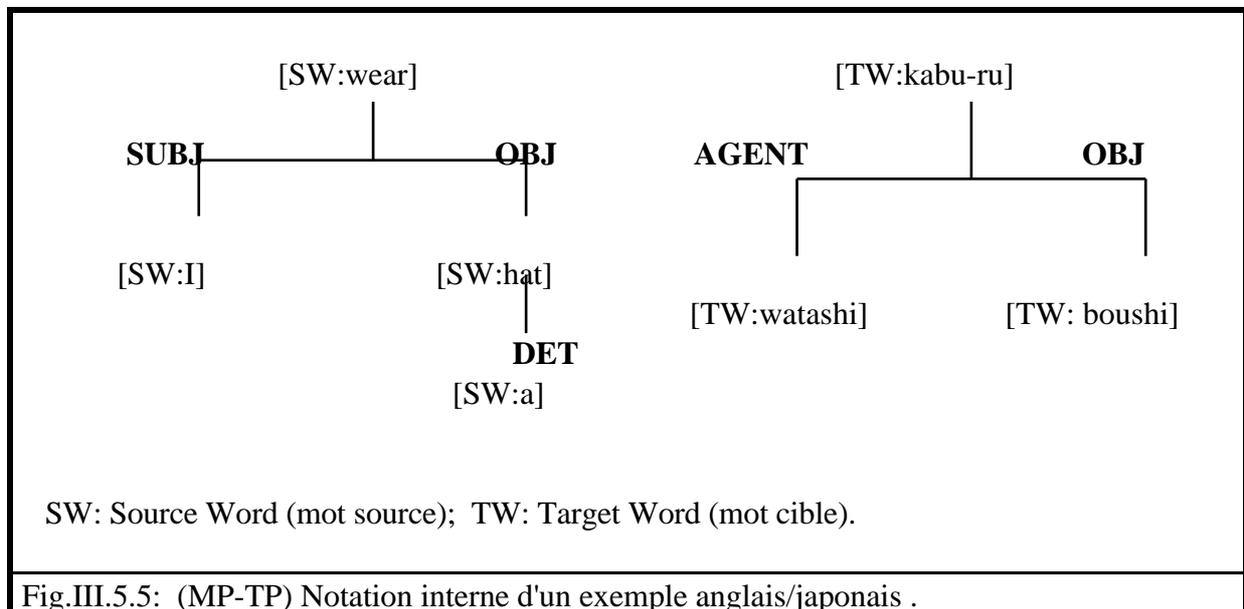
TP « Target Pattern »: correspondant à l'arbre à générer pour la cible.

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 noeuds sont connectés par des arcs nommés et le nom de l'arc dénote une relation. Voir [fig.III.5.5](#)



III.5.4.4 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.

III.5.4.5 La génération:

III.5.4.5.1 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.

III.5.4.5.2 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:

- ① stem : racine « yo »
- ② conjugation type : «5-dan»
- ③ kind of conjugational part : « ma »
- ④ 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 : negation]

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).

III.6. Le système TORJOMAN: [LABED 90],[LABED 92],[SIDHOM 94]

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.

III.6.1. 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).

III.6.2. 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 terme 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.

III.6.3. 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 .

III.6.3.1. 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 MiN DaAKa ?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 » : ?aLGaÿ .

⊕ 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 »: $Ma^C HaD \quad ?aL-BuHuT \quad Li^C uLuM \quad ?aL-?i^C LaAMiYa& \quad Wa \quad ?aL-?iTTiSaALaAT \quad ^c aN \quad Bu^C D.$

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

III.6.3.2. 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 .

III.6.3.3. 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.

III.6.4. Caractéristiques du système TORJOMAN :

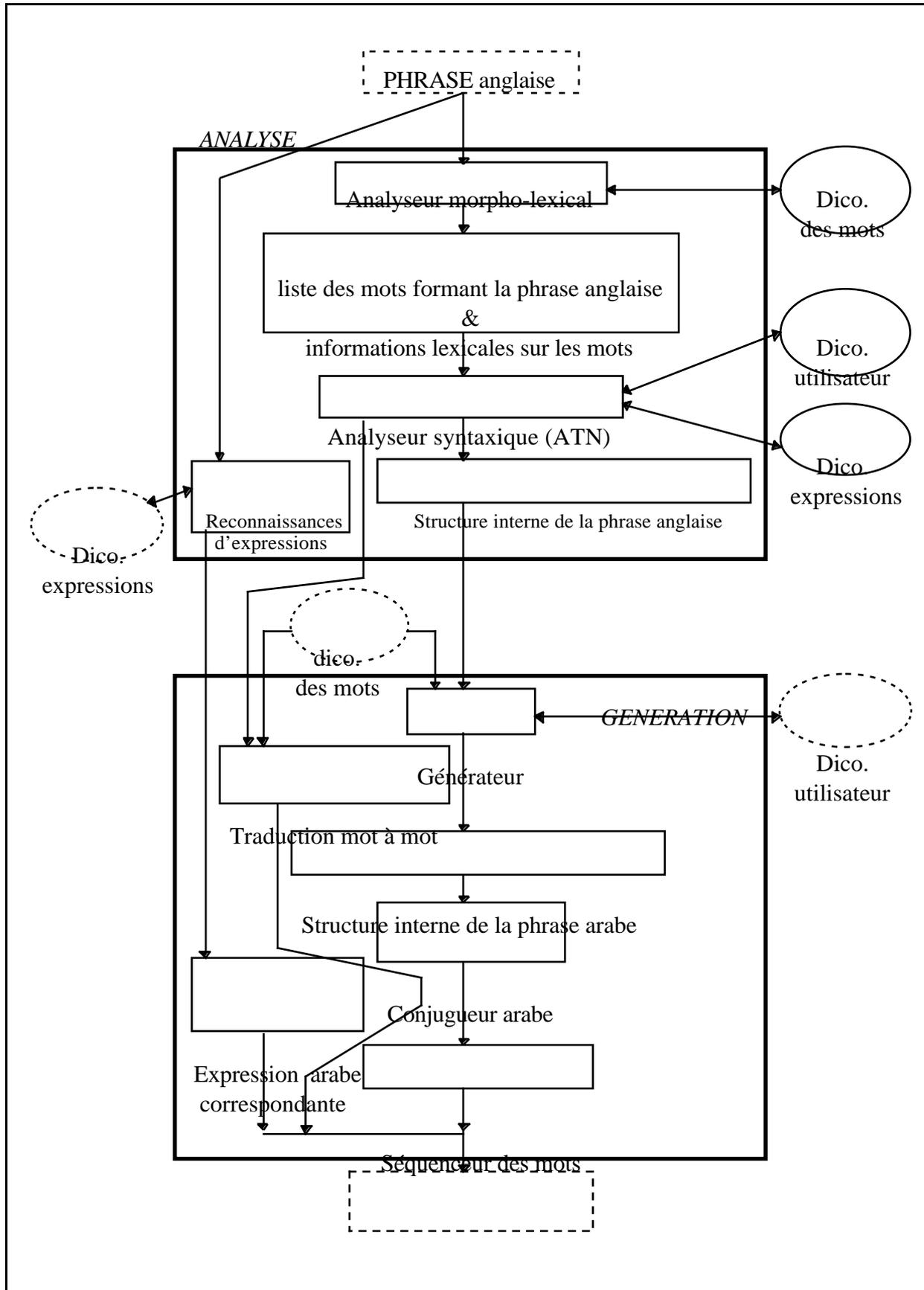
Torjoman dans des versions antérieures traduisait des phrases simples au mode déclaratif et la voix affirmative. Plusieurs améliorations ont permis l'évolution du système :

↯ 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.

III.6.5. Architecture du 'moteur' de traduction :



PHRASE ARABE

Fig.III.6.5: Architecture du système TORJOMAN.
--

III.6.6. 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,...

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 continu 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 95a]:

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 95b]:

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 95c]:

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

[AL-ABED 92]:

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 90]:

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 93]:

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

[AL-CHAY 88]:

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 95]:

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

[AMANO 89]:

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 91]:

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

[ARNOLD 87]:

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

[BALLON 94]:

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 90]:

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

[BILANGE 92]:

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

[BOLC 87]:

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

[BOUALEM 93]:

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 92]:

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

[BOURBEAU 85]:

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 87]:

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

[CHURCH 93]:

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

[CULLINFORD 87]:

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.

[COHEN 70]:

COHEN D., Essai d'une analyse automatique de l'arabe (1961-70), in *Etudes de linguistiques sémitique et arabe*, Paris: Mouton, 1970.

[DICHY 95]:

DICHY Joseph, Pour une lexicomatique de l'arabe: l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot, A paraître in *Lexicomaitique et dictionnaire* - Actes des journées de Lyon -, sept. 95.

[DICHY 90]:

DICHY Joseph, *L'écriture dans la représentation de la langue: la lettre et le mot en arabe*, Thèse de Doctorat d'Etat, Lyon 2 - Université Lumière, 1990.

[DICHY 89]:

DICHY Joseph et **HASSOUN M.O.**, *Simulation des modèles linguistiques et Enseignement Assistée par Ordinateur de l'arabe - Travaux SAMIA I*, Paris: Conseil International de la Langue Française, 1989.

[DORR 93]:

DORR Bonnie J., Interlingual machine translation a parameterized approach, *A.I N°63*, Elsevier Science Publishers, 1993. p.429-492.

[ELLUL 87]:

ELLUL COMMARC Aline, *Interface coopérative en L.N*, Thèse de Doctorat: I.A, ENST Paris, 1987. 162 p.

[FALZON 89]:

FALZON Pierre, *Ergonomie Cognitive du Dialogue*, Presse universitaire de Grenoble, 1989. 175 p.

[FARWELL 93]:

FARWELL D. (& **GUTHRIE L.**, **WILKS Y.**), Automatically creating lexical entries for ULTRA a multilingual MT system, *Machine Translation 8*, Kluwer Academic Publishers Netherlands, 1993. p.127-145.

[FORD 91]:

FORD Nigel, *Expert Systems and A.I*, Library Association Publishing London, 1991. 277 p.

[GAL 89]:

GAL A. (& **LAPALME G.**, **DIZIER P-St**), *Analyse et synthèse d'un langage - PROLOG*, Ed. EYROLLES, 1989. 135 p.

[GAZDAR 89]:

GAZDAR G. (& **MELLISH C.**), *Natural language processing in PROLOG*, Ed. Addison-Wesley Pub Company, 1989. 386 p.

[GROSS 95]:

GROSS Gaston, Une sémantique nouvelle de la TA- les classes d'objets, *La tribune des industries de la langue et de l'information électronique N° 17-18-19*, OFIL, 1995. p. 16-19.

[HASSOUN 87]:

HASSOUN Mohamed, *Système d'analyse morphologique automatique de la langue arabe*, Thèse de doctorat, Université Claude Bernard, Lyon 1, 1987. 142 p.

[HATON 91]:

HATON J-P (& al.), *Le raisonnement en I.A*, InterEditions, 1991. 480 p.

[JOHNSON 87]:

JOHNSON Roderick L. (& **WHITELOCK Peter**), MT as an Expert task, *Machine translation theoretical and methodological issues*, University Press of Combridge, 1987. p. 136-144.

[KABBAJ 91]:

KABBAJ A., *Intelligence artificielle en LISP et PROLOG*, Ed. MASSON, 1991. p. 257-322.

[KITREDGE 87]:

KITREDGE Richard I., The significance of sublanguage for automatic translation, *Machine translation theoretical and methodological issues*, University Press of Combridge, 1987. p. 59-67.

[KOUNIALI 93]:

KOUNIALI Samy Habib, *Désambiguïsation de groupes nominaux complexes en conformité avec les connaissances du domaine (Application à la traduction Automatique)*, Thèse de doctorat à l'INP de Lorraine, 23 septembre 1993. 185 p.

[LANGE 95]:

LANGE J-M, La TAO des dernières aux prochaines années, *La tribune des industries de la langue et de l'information électronique N° 17-18-19*, OFIL, 1995. p. 20-24.

[LAROUK 94]:

LAROUK OMAR, *Extraction des connaissances à partir de documents textuels: traitement automatique de la coordination (connecteurs et ponctuation)*, Thèse de doctorat, UCB Lyon1, Fev. 1994.

[LUZZATI 95]:

LUZZATI Daniel, *Dialogue Verbal Homme-machine*, Ed. MASSON, 1995. 197 p.

[LYTINEN 87]:

LYTINEN Steven L., Integrating syntax and semantics, *Machine translation theoretical and methodological issues*, University Press of Combridge, 1987. p. 302-316.

[MCDONALD 87]:

MCDONALD David D., Natural language generation: Complexities and techniques, *Machine translation theoretical and methodological issues*, University Press of Combridge, 1987. p.192-224.

[MELBY 87]:

MELBY Alan, On human-machine interaction in translation , *Machine translation theoretical and methodological issues*, University Press of Combridge, 1987. p. 145-154.

[METZGER 88]:

METZGER J-Paul, *Syntagmes nominaux et information textuelle- Reconnaissance automatique et représentation*, Thèse de doctorat, Université Claude Bernard, Lyon 1, 1988.324 p.

[MILLER 90]:

MILLER Philip (& **TORRIS Thérèse**), *Formalismes syntaxiques pour le traitement automatique du langage naturel*, Edition HERMES, 1990. 359 p.

[MONEIMNE 89]:

MONEIMNE Walid, *TAO vers l'arabe - Spécification d'une génération standard de l'arabe - Réalisation d'un prototype anglais-arabe à partir d'un analyseur existant*, Thèse de doctorat, Université Joseph Fourier, Grenoble 1, 14 juin 1989. 244 p.

[NAGAO 87]:

NAGAO Makoto, Role of structural transformation in a MT system, *Machine translation theoretical and methodological issues*, University Press of Combridge, 1987. p. 262-277.

[NIRENBURG 87a]:

NIRENBURG Sergei (& **RASKIN V.**, **TUCKER Allen B.**), The structure of interlingua in Translator, *Machine translation theoretical and methodological issues*, University Press of Combridge, 1987. p. 90-113.

[NIRENBURG 87b]:

NIRENBURG Sergei, Knowledge and choices in MT, *Machine translation theoretical and methodological issues*, University Press of Combridge, 1987. p. 1-21.

[PEREIRA 80]:

PEREIRA Fernando C.N. & **WARREN David H.D.**, Definite Clause Grammars for language analysis- A survey of the formalism and comparison with Augmented Transition Networks, *A.I n° 13* , 1980, p.213-278.

[PITRAT 85]:

PITRAT Jacques , *Textes Ordinateurs et Compréhension*, Ed. EYROLLES, 1985. 201 p.

[RADWAN 94]:

RADWAN Khaled, *Vers l'accès multilingue en langage naturel aux bases de données textuelles*, Thèse de doctorat , Université de Paris-Sud, Centre d'ORSAY, 11/02/1994. 243 p.

[RASKIN 87]:

RASKIN Victor, Linguistics and natural language processing, *Machine translation theoretical and methodological issues*, University Press of Combridge, 1987. p. 42-58.

[RICHARD 95]:

RICHARD Jean-Luc, *STELO LINGVA: un système de TA centré sur l'Espéranto - Phase0: la traduction Espéranto-Français*, Mémoire de DEA, INSA de Lyon, Sep. 1995. 58 p.

[ROMAN 90]:

ROMAN André, *Grammaire de l'Arabe*, Paris: P.U.F. (coll. 'Que sais-je?'), 1990.

[SABAH 90a]:

SABAH Gérard, *L'intelligence artificielle et le langage - Volume 1: représentation des connaissances*, Ed. HERMES, 1990. 352 p.

[SABAH 90b]:

SABAH Gérard, *L'intelligence Artificielle et le langage Volume 2: Processus de compréhension*, Edition HERMES, 1990. 411 p.

[SHIEBERT 88]:

SHIEBERT Y., *Lexicalized Grammars: on application to tree adjoining grammars*, acte de colloque: 12th colling Budapest, 1988.

[SIDHOM 94]:

SIDHOM Sahbi, *Le système TORJOMAN: traducteur assistée par ordinateur de l'anglais vers l'arabe*, Nov.1993-94, IRSIT Tunisie, p. 1-12 , Rapport d'activités sur le système TORJOMAN.

[SOMERS 93]:

SOMERS Harold L., Current research in MT, *M.T N°7*, Kluwer Academic Publishers Netherlands, 1993. p. 231-246.

[TUCKER 87]:

TUCKER Allen B., Current strategies in MT research and development, *Machine translation theoretical and methodological issues*, University Press of Combridge, 1987. p. 22-41.

[VAUQUOIS 75]:

VAUQUOIS Bernard, *La traduction automatique à Grenoble*, DUNOD, 1975. 179 p.

[VEGA 95]:

VEGA J., AlethTrad outil et service, *La tribune des industries de la langue et de l'information électronique N° 17-18-19*, OFIL, 1995. p. 24-27.

[VEILLON 70]:

VEILLON Gerard, *Modèles et algorithmiques pour la traduction automatique*, Thèse de Doctorat, Faculté des sciences de l'université de Grenoble, 1970. 156 p.

[WHITE 87]:

WHITE John S., The research environnement in the METAL project, *Machine translation theoretical and methodological issues*, University Press of Combridge, 1987. p. 225-246.

[WILKS 94]:

WILKS Y., Developments in MT research in the US, *Aslib Proceedings vol.46 N°4*, April 1994. p.111-116.

[WILLS 93]:

WILLS wolfram, Basic concepts of MT, *Meta XXXVII.3*,1993. p. 403-413.

[ZAHARIN 88]:

ZAHARIN Yossoff , Towards an analyser (parser) in a machine translation system based on ideas from expert systems, *Comput. Intelligence 4*, 1988. p. 180-191.