I - HISTORIQUE DU PROJET

Nos recherches ont débuté il y a bientôt quatre ans. Ce projet fut d'abord intégré aux activités générales de la Faculté des Lettres de l'Université de Montréal; maintenant, il dépend directement du vice-recteur à la recherche de la même université.

Nous sommes particulièrement reconnaissants à l'égard du premier directeur de ce projet, M. Guy Rondeau, qui obtint une subvention du Conseil National de la Recherche (laquelle est renouvelée annuellement) et envers la direction du projet C.E.T.A. à Grenoble pour avoir favorisé la formation de plusieurs d'entre nous.

Notre groupe est maintenant composé de quatre linguistes et de quatre ingénieurs-informaticiens ou programmeurs dont M. Alain Colmerauer qui agit comme directeur depuis peu.

Les travaux que nous conduisons et dont nous allons vous entretenir concernent uniquement ceux qui sont liés au projet de la traduction automatique.

Le document que nous vous avons présenté ne fait pas état non plus des recherches antérieures, notamment celles qui portèrent sur l'utilisation de divers modèles de reconnaissance applicables aux langues naturelles.

En ce qui concerne ces recherches, des rapports périodiques continuent de paraître et le dernier d'entre eux ajoute considérablement aux notes trop succintes qui vont suivre.

II - GRAMMAIRES-W

Nous ne donnerons ni une histoire ni une description formelle du système-W; nos collègues informaticiens le font dans une communication au Congrès de l'A.C.M., San Francisco, Août 1969 [DE CHASTELLIER et COLMERAUER (1969)]. Nous voulons plutôt expliquer rapidement l'usage linguistique que nous faisons des caractéristiques de ce système.

Celui-ci consiste en un parseur-interpréteur (P.I.) et un synthétiseur, tous deux de puissance transformationnelle. Le synthétiseur est l'inverse du P.I. à quelques détails près, et nous ne considerons que le P.I. dans ce qui suit. Les entrées de dictionnaire sont écrites dans le même format et avec le même statut que les autres règles, et jusqu'ici il n'y a pas de traitement séparé pour le dictionnaire [cf. section V].

Le système est essentiellement destiné à traîter des chaînes, quoiqu'il permette aussi, comme nous le verrons, de traiter des arbres. La donnée d'entrée pour le P.I. est constituée d'une chaîne et d'une grammaire; la sortie est une (ou des) chaîne(s) "interprétée(s)". La différence entre chaîne et chaîne interprétée est la suivante: la première est une séquence non concaténée, que l'on peut concaténer par la suite en lui appliquant des règles appropriées; la seconde a subi la concaténation et forme un symbole complexe unique. La conséquence, parfois gênante, en est que le P.I. ne peut traiter une chaîne interprétée que dans sa totalité. Par exemple, si on a appliqué une règle:

$$\text{THE}\frown\text{MAN} + \text{SPOKE} \longrightarrow \text{THE}\frown\text{MAN}\frown\text{SPOKE}$$

il sera impossible d'appliquer des règles subséquentes à SPOKE, à moins que la règle ne spécifie THE⌒MAN aussi.

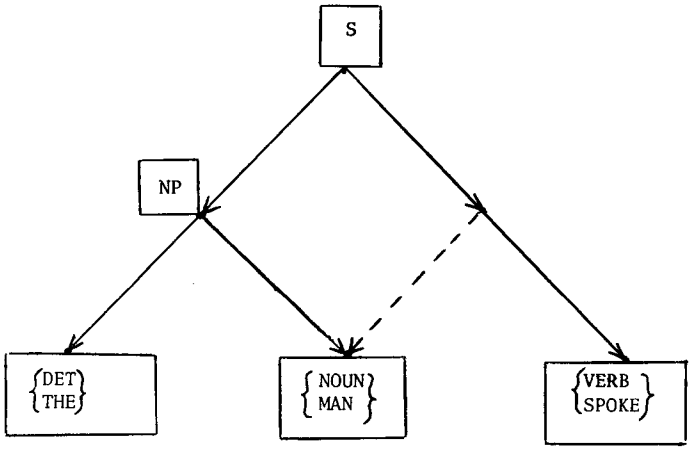
Les structures d'arbre peuvent être indiquées par des chaînes parenthésées. De même on peut introduire des étiquettes de noeuds ou de fonctions dans les chaînes; il n'y a pas d'indices souscrits.

Dans la description formelle, les chaînes interprétées finales sont appelées "chaînes axiomatiques". (Le synthétiseur commence sa dérivation à partir de ces "chaînes axiomatiques".) Dans notre système de traduction, la donnée de P.I. comporte une chaîne anglaise [cf. section VI]; les "chaînes axiomatiques" sont des chaînes de notre langage pivot [cf. section IV]; la sortie du synthétiseur est une traduction en un français "restreint" [cf. section VII]. Le P.I. et le synthétiseur peuvent être enchaînés pour passage direct de l'anglais au français.

Une grammaire W est faite de deux parties, i.e. deux ensembles de règles disjoints, qui sont décrits ci-dessous. Le système applique chacune des deux parties l'une après l'autre, à chaque étape du traitement. Cette alternance continuelle est peut être le trait le plus inhabituel de la grammaire W, pour un linguiste, et les commençants prennent en général quelques semaines pour s'y faire.

Le P.I. avance de gauche à droite sur la chaîne d'entrée, ajoutant un symbole à la fois (séparé par des blancs de part et d'autre) au segment déjà interprété. Pour chaque segment, il établit un réseau de noeuds étiquetés, que nous pouvons représenter comme un demi treillis [cf. fig. 1].

Fig. 1



Meta-rules:

1. S = NP VERB
2. NP = DET NOUN
3. DET = THE
4. NOUN = MAN
5. VERB = SPOKE

--> :this arc completes the graph but is not used in the PS structure specified by the meta-rules.

Une des parties de la grammaire est constituée de "métarègles". Celles-ci sont C.F. Le système les utilise pour étiqueter certains noeuds du réseau. Les règles unaires attachent plusieurs étiquettes au même noeud. Nous pouvons considérer que les autres fabriquent des arbres de structures constituants, si nous écrivons la grammaire dans ce sens.

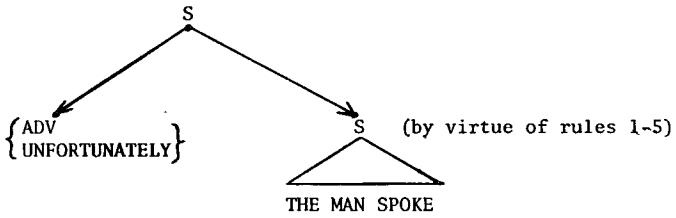
Une chaîne ainsi traitée peut désormais être décrite par le profil d'une section à travers le réseau que la métagrammaire lui a attaché. Pour cela, il fait énoncer, de gauche à droite, les noms des noeuds que le profil traverse - et qui doivent donc avoir été introduits par les métarègles - Ainsi la chaîne de la fig. 1 peut se décrire par "DETERMINER MAN SPOKE" ou "NP SPOKE", ou tout autre profil correct. Enfin, tout un "arbre" peut être dénoté simplement par le nom de son noeud supérieur. [fig. 2]

Fig. 2

Additional meta-rules:

6. S ⟶ ADV S

7. ADV ⟶ UNFORTUNATELY



Sur chaque segment, le système effectue toutes les analyses possibles selon les métarègles. Comme dans d'autres algorithmes (par exemple celui de Cocke) ceci produit des analyses de sous-chaînes qui s'avèreront abortives avant que le traitement de toute la chaîne d'entrée soit achevé. Il faut admettre que cette stratégie charge la mémoire de l'ordinateur, et pour le moment nos chaînes sont limitées à environ 30 symboles. [mais voir Section VIII]

La description par profils s'est avérée très utile pour l'écriture des autres règles de la grammaire, appelées "pseudorègles" (ce nom n'est pas très révélateur quant à l'usage que les linguistes font de ces règles).

Dans la partie transformationnelle, on peut donc écrire des "pseudorègles" qui, jointes aux métarègles, génèrent un langage de type 0. Il n'y a pas de restriction sur le nombre de symboles ni les types d'opérations dans ces règles. Comme un symbole peut nommer un arbre - grâce aux métarègles - le résultat pratique est un pouvoir considérable dans le traitement des arbres. En fait nous pensons que le système W pourrait être utilisé comme testeur de grammaires transformationnelles, moyennant l'introduction de certaines améliorations, par exemple ordre des règles.

La fig. 3 donne un exemple de grammaire W - linguistiquement élémentaire mais complète - ainsi qu'une sortie correspondante. Dans le P.I., les règles doivent se lire " < membre droit > est réécrit < membre gauche > ". Pour la clarté de l'exposé, nous avons quelque peu changé le format de la sortie machine réelle.

Fig. 3

(OPERATORS)

== : IS RE-WRITTEN

\+ : CONCATENATED WITH

$: END OF A RULE, OR OF A SET OF RULES WITH IDENTICAL LEFT-HAND MEMBERS

* : COMMENT BOUNDARY

* THE SYMBOLS USED FOR THE ABOVE OPERATORS ARE CHOSEN BY THE LINGUIST AND ARE INPUT WITH THE DATA *

(META-RULES)

1) X REPRESENTS ANY SINGLE TERMINAL SYMBOL

* THIS RULE IS INCORPORATED IN THE PROGRAM *

2) INTRANSITIVE-SENTENCE = VERB SUBJECT-(NOUN-PHRASE) . $

3) NOUN-PHRASE = ADJECTIVE NOUN

4) = DETERMINER NOUN-PHRASE $

5) ADJECTIVE = (ADJ X) $

6) NOUN = (NN X) $

7) DETERMINER = (A-REFERRED-TO) $

* AS DETERMINERS ARE A SMALL CLOSED CLASS IT IS A WORTHWILE SIMPLIFICATION MERELY TO LIST THEM . *

8) VERB = (VB X TENSE) $

9) TENSE = PAST $

INTRANSITIVE-SENTENCE * THIS IS THE DECLARATION OF THE SUMMIT SYMBOL WHICH IS TO TERMINATE THE PARSING PROCESS PROVIDED THE WHOLE INPUT STRING IS INCLUDED UNDER IT *

* NOTE TOO THAT A SYMBOL IS DEFINED IN THE PROGRAM AS AN UNBROKEN STRING OF CHARACTERS BETWEEN TWO BLANKS, SO THAT A HYPHENATED COMPOUND IS ONE SYMBOL *

(PSEUDO-RULES)

10) (ADJ HAIR-LORN) = BALD $

11) (NN MALE-HUMAN) = MAN $

12) (VB SPEAK PAST) = SPOKE $

13) (A-REFERRED-TO) = THE $

* THERE IS NO NEED OF A PART-OF SPEECH MARKER FOR THE DETERMINER, BECAUSE IT IS ALLOCATED SPECIFICALLY TO ITS SYNTACTIC CATEGORY BY RULE 7 *

14) (NOUN-PHRASE) = NOUN-PHRASE $

* A TYPICAL RULE FOR INSERTING PARENTHESES, WHICH THEN HAVE THE FORMAL STATUS OF TERMINAL SYMBOLS *

15) ADJECTIVE NOUN = ADJECTIVE + NOUN $

16) DETERMINER NOUN-PHRASE = DETERMINER + NOUN-PHRASE $

* CONCATENATION RULES LIKE 15 THROUGH 17 ARE REQUIRED FOR REASONS TO DO WITH THE FORMALISM OF THE W-SYSTEM *

17) SUBJECT-(NOUN-PHRASE) VERB .

= (NOUN-PHRASE) + VERB + . $

* A TYPICAL RULE FOR INSERTING A LABEL OF SYNTACTIC FUNCTION *

18) VERB SUBJECT-(NOUN-PHRASE) .

= SUBJECT-(NOUN-PHRASE) VERB . $

* TYPICAL RULE FOR TRANSPOSING A SUB-TREE.
ACTUALLY RULES 17 AND 18 COULD BE COMBINED
INTO ONE RULE *

(INTERPRETATIONS)

* ABORTIVE INTERPRETATIONS HAVE BEEN SUPPRESSED FOR BREVITY *

* INPUT STRING: THE BALD MAN SPOKE . *

1) (A-REFERRED-TO) = THE * BY RULE 13 *

2) (ADJ HAIR-LORN) = BALD * BY RULE 10 *

3) (NN MALE-HUMAN) = MAN * BY RULE 11 *

4) (ADJ HAIR-LORN) (NN MALE-HUMAN) = LINES 2 + 3
* BY RULES 5, 6, 15 *

5) (A-REFERRED-TO) (ADJ HAIR-LORN) (NN MALE-HUMAN)
= LINES 1 + 4 * BY RULES 7, 5, 6, 3,
16 *

6) ((A-REFERRED-TO) (ADJ HAIR-LORN) (NN MALE-HUMAN))
= LINE 5 * BY RULES 7, 5, 6, 3, 4,
14 *

7) (VB SPEAK PAST) = SPOKE * BY RULE 12 *

8) SUBJECT-((A-REFERRED-TO) (ADJ HAIR-LORN)
(NN MALE-HUMAN))) (VB SPEAK PAST) .
= LINES 6 + 7
* BY RULES 7, 5, 6, 3, 4, 9, 8, 17 *

9) (VB SPEAK PAST) SUBJECT-((A-REFERRED-TO)
(ADJ HAIR-LORN) (NN MALE-HUMAN)) .
= LINE 8
* BY RULES 7, 5, 6, 3, 4, 9, 8, 18 *
* LINE 9 IS THE SOLE ULTIMATE INTERPRETATION - IN FORMAL TERMS, THE SOLE AXIOMATIC STRING - SINCE THE WHOLE INPUT STRING IS DERIVABLE FROM IT AND IT IS DOMINATED, ACCORDING TO THE METAGRAMMAR, BY 'INTRANSITIVE-SENTENCE' WHICH HAS BEEN DECLARED AS THE NAME OF THE TERMINATING OPERATOR *

III - ESQUISSE D'UN MODELE DE TRADUCTION

0. La disponibilité du système W, si puissant, sous certains rapports, a poussé certains d'entre nous à la réflexion théorique. Il nous fallait en effet faire des choix parmi les possibilités offertes pour le traitement des données linguistiques.

On peut se représenter idéalement la "faculté de langage" d'un homme comme une machine non déterministique qui a pour fonction d'effectuer une correspondance entre certaines chaînes sonores et certaines représentations sémantiques conçues comme une collection de relations choisies (par une opération que nous pouvons qualifier d' "abstraction") parmi celles qui existeraient entre des éléments de la perception.

La traduction idéale, comme chacun sait, consiste à "comprendre" un texte dans une langue - c'est-à-dire à construire là où les représentations sémantiques correspondantes - et à "parler" dans une autre langue - c'est-à-dire effectuer dans l'autre langue les opérations conduisant de la représentation sémantique à la chaîne sonore. Idéalement, la seconde phase devrait être fortement conditionné par la première, et nous devrons examiner les points de correspondance.

1. La faculté de langage

Il est utile d'imaginer les structures sémantiques comme des graphes (qui n'ont a priori aucune raison d'être des arborescences) dont les noeuds - ou les arêtes - sont étiquetés par des indices de référence et des noms de relations existant entre certains de ces indices. Le "problème" du locuteur est de représenter ces structures sous forme de chaînes parenthésées, c'est-à-dire, en première approximation sous forme de structures de constituants. Ce problème a deux aspects: d'une part, la représentation de la structure du graphe d'autre part la représentation sonore (ou graphique) des éléments "substantifs" (i.e. des étiquettes relationnelles de ce graphe).

La première partie est communément appelée syntaxe. Dans les termes employés ici, elle consiste à créer des structures de constituants pour représenter une partie des configurations du graphe sémantique. Nous nous représenterons cette partie de la faculté de langage comme une collection de modules opératoires, chacun effectuant une seule opération sous le contrôle de paramètres décrivant les structures sémantiques à transformer. [Voir HOFMANN

(1968) pour un exemple d'une telle opération]

La deuxième partie est le lexique. Celui-ci peut être vu comme une relation binaire dont les premiers arguments sont des paquets de traits sémantiques et les seconds arguments des matrices phonologiques (-Dans le cas d'un système opérant sur des textes écrits, le lexique sera une relation entre des paquets de traits sémantiques et des chaînes de caractères). La relation "lexique" n'est pas une fonction, en ce qu'un paquet de traits sémantiques donné peut avoir un correspondant phonique différent selon des circonstances paralinguistiques. Il est intéressant d'introduire là des paramètres d' "attitude" , "style" etc. [voir section V] Mentionnons en passant qu'il est séduisant (et après tout raisonnable) de faire l'hypothèse suivante: la forme des divers modules syntaxiques et l'ensemble de tous les traits sémantiques semblent être universels. Par contre les valeurs des paramètres de contrôle des modules syntaxiques, ainsi que la relation "lexique" avec tous ses paramètres, semblent être acquis par éducation dans une société donnée.

2. Une traduction consistera donc en deux fois deux opérations. Premièrement, des formes phoniques P_1 - ou écrites - seront identifiées et appliquées sur des paquets de traits sémantiques S par la relation "lexique 1" tandis que les structures de constituants seront analysées et appliquées sur un graphe - étiquetées par les S - par l'opération des modules syntaxiques (les paramètres ayant les valeurs P_1).
Deuxièmement les graphes ainsi obtenus seront traités par les mêmes modules syntaxiques, dont les paramètres auront pris les valeurs P_2 correspondant à la langue cible, tandis que les étiquettes S seront appliquées sur des chaînes de caractères ou des matrices phonologiques P_2 par la relation "lexique 2".

3. Cette représentation rationnalise (il n'est pas possible de parler de véritable "justification") l'approche que nous avons adoptée en ce qui concerne le traitement automatique du problème de traduction, telle qu'elle est présentée dans les sections suivantes.

Notons un fait important. Les paramètres syntaxiques ou lexicaux acquis par éducation dans une certaine société sont fortement variables à l'intérieur même d'une langue donnée (variations "dialectales", "stylistiques" etc.) L'attitude et la sensibilité du traducteur envers ces variations peut différer énormément. L'idéal serait que la traduction respecte toutes les nuances. Dans le cas de la traduction automatique, toutefois, ceci impliquerait une "analyse culturelle" qui reste à faire. Voici une décision possible: on choisira d'accepter autant de structures possi-

bles dans la langue source, c'est-à-dire qu'on permettra aux paramètres la plus large variation compatible avec l'intelligibilité (élargir le domaine de variation d'un paramètre de contrôle diminue naturellement l'information apportée par l'opération correspondante). On essaiera tout de même de corréler autant que possible les valeurs des paramètres - surtout dans le lexique - avec des niveaux de style, etc. Du côté de la langue cible, on se contentera pour un temps de transmettre l'information nécessaire, c'est-à-dire qu'on restreindra fortement la variation des paramètres autour de la valeur correspondant approximativement au "standard" du langage. (On pourra même aller au delà en restreignant chaque structure à une seule expression dans la langue cible.

IV - LE LANGAGE PIVOT

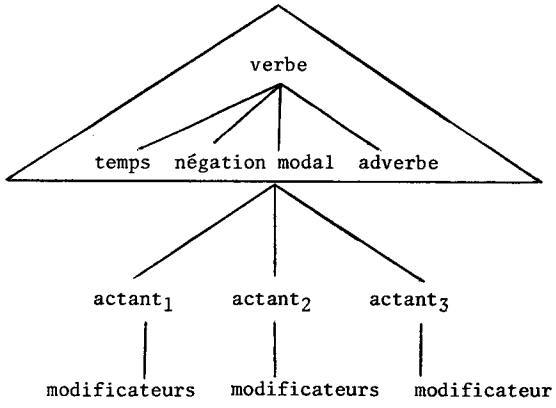
Le langage pivot est un langage formel apte à définir des relations sémantiques. Les éléments qui le composent sont des mots - appelés z-mots dans notre système - qui correspondent d'une façon univoque à des configurations sémantiques d'un type particulier (restreintes dans notre système actuel à la paire anglais-français); ces mots ne sont donc pas ambigus et n'ont pas de synonymes. L'ordre canonique dans lequel ces z-mots sont disposés indique les relations sémantiques qu'il entretiennent.

Par rapport au modèle de traduction, une chaîne du langage pivot ne fournit qu'une représentation sémantique pour chaque chaîne de lexèmes de la langue source et pour une ou plusieurs chaînes de lexèmes dans la langue cible. Tel qu'utilisé dans notre système, le langage pivot fournit des chaînes qui constituent la sortie de l'analyseur de l'anglais et deviennent en même temps l'entrée pour le générateur du français. Plus explicitement, l'analyseur transforme une suite de mots et de signes de ponctuation de l'anglais en autant de chaînes canoniques du langage pivot qu'il y a de sens différents attribués à cette suite. Le générateur va traiter cette chaîne canonique et la transformer à son tour en une ou plusieurs suites de lexèmes du français dont une au moins doit correspondre à la configuration sémantique exprimée dans la chaîne du langage pivot.

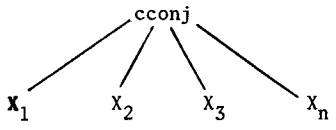
Le z-mot auquel correspond un ensemble défini de lexèmes appartient au lexique. On obtient autant de z-mots pour un lexème que celui-ci a de sens différents d'ambiguité; par contre, un seul z-mot recouvre toute une classe de lexèmes synonymes. Ces z-mots ne constituent pas des traductions françaises de mots anglais mais plutôt des entités abstraites qui recouvrent une ou plusieurs configurations sémantiques particulières qui prendront dans différentes

langues naturelles différentes configurations graphémiques ou phonémiques.

Les relations qu'entretiennent les éléments du langage pivot sont exprimées par des structures de dépendance. Celles-ci ont été imaginées tout d'abord par Tesnière, et l'usage que nous en faisons est inspiré des études effectuées à Grenoble; toutefois la forme dans laquelle nous les utilisons est conditionnée par nos besoins particuliers. Le tableau suivant illustre la structure de dépendance caractéristique de notre modèle.



Une autre structure consiste dans la coordination d'éléments d'une même classe (par exemple des adjectifs); elle correspond à:



L'adaptation des structures de dépendance se fait naturellement à mesure que la grammaire devient plus complexe ou que le formalisme reçoit des modifications. Par exemple, nous voulons introduire sous peu des spécificateurs de phrases comme "interrogatif", "impératif", etc. et inclure dans cette classe d'opérateurs ceux du temps, de la négation, du modal et de l'adverbe.

La sortie des chaînes de ce modèle de dépendance correspond actuellement à une représentation linéaire des éléments. Le gouverneur précède la chaîne gouvernée elle-même entre parenthèses. Dans le cas où la fonction sémantique révélée dans chacune des chaînes gouvernées par un même gouverneur n'est pas la même, comme cela se produit pour les premier, deuxième, troisième actants, on ajoute des étiquettes pour indiquer le type de dépendance dans chacun des cas. *L'étiquette "epi-" s'emploie pour marquer des chaînes qui modifient la chaîne qui les gouverne. Par exemple, les relatives sont précédées de cette étiquette parce qu'elles modifient le groupe nominal qui les gouverne.

V - LEXICOGRAPHIE

0. Jusqu'à la rédaction de cet article, notre travail lexicographique s'est surtout occupé de la "conversion" des mots anglais en lexèmes de notre langage intermédiaire (pivot).

La figure 1 illustre une entrée de dictionnaire pour un verbe anglais. La ligne 2 de l'exemple présente ce que nous appelons l'entrée primaire: elle contient toute l'information notée par notre lexicographe, y compris une citation. La citation est encadrée par des astérisques, et traitée comme un "commentaire" par les programmes de traitement. Au-dessus de chaque symbole de l'entrée primaire apparaît la variable lexicale dont le symbole est une valeur. Dans le système W que nous utilisons [voir section II] ces variables sont définies dans les "métarègles", soit par listes de valeurs, soit par un schéma de la forme:

ZWORD = Z X / X représente n'importe quelle valeur

Les entrées de dictionnaire proprement dites sont écrites dans le format général des "pseudorègles" de la grammaire-W, et ont le même statut que les autres règles de ce type. En conséquence elles peuvent être soumises directement aux règles

* On pourrait se contenter de l'ordre des chaînes.

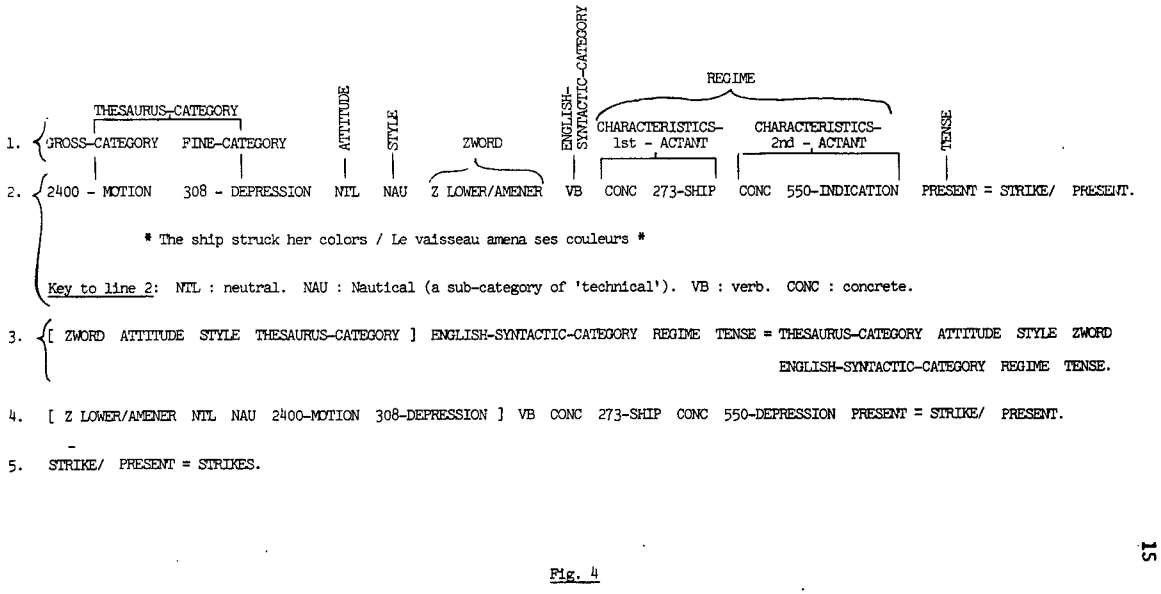

THESAURUS-CATEGORY
REGIME
1. GROSS-CATEGORY FINE-CATEGORY ATTITUDE STYLE ZWORD ENGLISH-SYNTACTIC-CATEGORY CHARACTERISTICS- 1st - ACTANT CHARACTERISTICS- 2nd - ACTANT TENSE
2. 2400 - MOTION 308 - DEPRESSION NTL NAU Z LOWER/AMENER VB CONC 273-SHIP CONC 550-INDICATION PRESENT = STRIKE/ PRESENT.
* The ship struck her colors / Le vaisseau amena ses couleurs *
Key to line 2: NTL : neutral. NAU : Nautical (a sub-category of 'technical'). VB : verb. CONC : concrete.
3. [ZWORD ATTITUDE STYLE THESAURUS-CATEGORY] ENGLISH-SYNTACTIC-CATEGORY REGIME TENSE = THESAURUS-CATEGORY ATTITUDE STYLE ZWORD ENGLISH-SYNTACTIC-CATEGORY REGIME TENSE.
4. [Z LOWER/AMENER NTL NAU 2400-MOTION 308-DEPRESSION] VB CONC 273-SHIP CONC 550-DEPRESSION PRESENT = STRIKE/ PRESENT.
5. STRIKE/ PRESENT = STRIKES.

Fig. 4

d'interprétation de la grammaire, et peu importe le format choisi pour les entrées primaires, tant qu'elles satisfont au format général des grammaires-W. En effet on peut écrire facilement les règles nécessaires pour les transcrire dans un format compatible avec la partie "syntaxique" de la grammaire. Ainsi on peut introduire beaucoup de détails descriptifs dans les entrées de dictionnaire même si nous n'en avons pas l'usage pour le moment. Malheureusement, chaque règle de traitement du dictionnaire accroît la grammaire et par conséquent le temps de traitement, parce que le programme d'application de la grammaire W essaie chaque règle à chaque phase du traitement. Il est évident qu'il nous faudra écrire sous peu un programme particulier de consultation du dictionnaire, même si nous pouvons nous en tirer pour le moment en enchaînant deux grammaires-W, dont la première ne se compose que de règles de ce niveau.

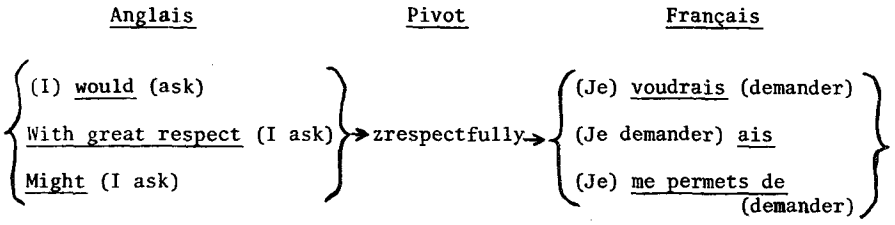
La ligne 3 dans la figure est la règle qui réécrit l'entrée primaire dans le format requis à présent dans notre grammaire. La ligne 4 est le résultat de l'application de cette règle. Toutes les valeurs encadrées par des crochets sont "persistantes" - c'est-à-dire sont conservées dans le pivot - Les lexèmes du pivot sont donc des symboles complexes limités par des crochets.

La ligne 5 illustre une règle "morphologique", qui explicite le temps du verbe et le met sous la forme RACINE/TEMPS correspondant à la ligne 4. Cette règle fait également partie des "pseudorègles", et n'a pas de statut particulier.

Nous allons à présent examiner l'importance linguistique des diverses variables.

1. ZWORD

> Nous avons donné il y a quelque temps dans un article [Harris (1968)] une description formelle des z-mots. Nous considérions qu'un élément de ce vocabulaire de z-mots était un symbole pour une variable sémantique dont les valeurs étaient des sens de lexèmes anglais et français. En fait, nous posions à la création d'un z-mot la condition qu'il existait dans chaque langue au moins un lexème dont un sens appartiendrait au domaine de variation de la variable à représenter par ce z-mot. Les valeurs peuvent être des racines lexicales sujettes à déclinaison et affixation, ou des formes complètes de surface, ou des morphèmes de profondeur sujets à des transformations de la grammaire. Exemple:



Tous les éléments de l'ensemble de sens anglais sont synonymes; les éléments de l'ensemble français sont synonymes aussi et traduisent l'ensemble anglais. Ainsi un z-mot était considéré comme une étiquette pour une fonction de traduction dont les arguments étaient un ensemble de synonymes en anglais et un en français. Il n'est peut être pas complètement inutile ici de mettre en garde contre des présomptions d'universalité: jusqu'ici notre recherche a été strictement limitée à l'anglais et au français. Toutefois, le modèle peut théoriquement être étendu pour comprendre plus de deux langues et construire les expansions complexes de ce qu'Andreyev a appelé "anneaux de traduction" [ANDREYEV (1965)].

Depuis la publication de cet article, rien n'a changé dans le statut formel des z-mots; en ce qui concerne le traitement, toutefois beaucoup de changements ont eu et ont encore lieu. Pour garder nos chances des respecter les délais de notre contrat, nous avons réduit la portée de notre modèle. Pour le langage source, il nous faut nous préparer à accepter autant que possible du vocabulaire anglais - si nous ne voulons pas restreindre notre système à une classe étroite de textes techniques - Par contre il nous suffira pour le moment d'utiliser à la synthèse une portion bien plus réduite du vocabulaire français. Dans l'exemple ci-dessus, cette restriction temporaire nous permet de nous dispenser de deux des trois réalisations françaises de "zrespectfully". Nous avons de plus accepté une hypothèse supplémentaire: qu'il existe un vocabulaire "de base" du français, servant à exprimer toutes idées communes, et qu'il suffit d'enrichir de micro-glossaires techniques pour permettre une traduction des textes anglais. Autrement dit, un vocabulaire correspondant au Basic English. Nous avons la chance de posséder une bonne approximation d'un tel vocabulaire dans les deux livres Le Français fondamental (F.f.) [GOUGENHEIM (s.d.)]. Jusqu'ici la technique la plus rapide que nous ayons trouvée pour la compilation du dictionnaire est la "traduction à rebours" du F.f. vers l'anglais.

Nous avons considéré diverses manières d'étiqueter la fonction ZWORD. Les semoglyphes d'Andreyev devaient être numériques. Mais nous n'avons pas accepté l'idée d'un langage pivot qui ne serait pas directement lisible par les linguistes qui l'utilisent. Au début, nous utilisions des mots français que nous distinguions du vocabulaire du français proprement dit par le préfixe "z" (D'où le terme "z-mot"). Récemment, nous avons adopté un étiquetage dont les symboles sont plus longs, mais plus explicites. Ayant établi un "ensemble de synonymes" anglais, nous attachons une étiquette anglaise à cet ensemble. Cela est aisé, puisque nous avons un bon dictionnaire anglais des synonymes qui donne une étiquette convenable pour chaque collection de synonymes [LEWIS (1961)]. Du côté français nous n'avons pas de collections de synonymes pour le moment, mais un seul lexème F.f. . Le z-mot est composé par composition de l'étiquette anglaise avec le lexème F.f. Par exemple, certains usages de "give up", "renounce", "surrender", "abandon", etc., sont tous couverts par l'étiquette RELINQUISH dans le dictionnaire de synonymes, et tous traduisibles par 'abandonner', lexème F.f. . Nous composons donc un z-mot "zrelinquish/abandonner", pour marquer la 'variable de traduction' dont les valeurs sont les mots cités ci-dessus. Par la suite d'autres valeurs, synonymes de "abandonner" seraient introduites.

Nous avons conscience d'un certain danger: en opérant par "traduction à rebours" à partir des lexèmes F.f. nous pourrions avoir des difficultés si le F.f. contient déjà des synonymes: une collection de synonymes anglais pourrait se trouver marquée de plusieurs z-mots différents; cela créerait de l'ambiguïté dans le langage pivot, ce que nous désirons vivement éviter. Si une telle ambiguïté apparaît, nous aurons à raffiner la notion de z-mot pour l'éliminer. Mais nous avons déjà introduit des raffinements, puisque nous conservons dans le pivot, outre les z-mots, les éléments étudiés plus bas, soit: SEMANTIC PARAMETERS, STYLE, ATTITUDE, THESAURUS CATEGORY.

Par la manière dont les z-mots sont produits, nous obtenons une définition opérationnelle de la valeur cognitive des lexèmes du pivot. Chaque collection de sens de lexèmes constitue ce que Sparck Jones, dans son ouvrage sur les synonymes anglais, appelle un "row"; nous considérons que nous avons étendu son modèle à la traduction lexicale ou, disons-nous, à la "synonymie interlinguale" [SPARCK JONES (1965)].

2. SEMANTIC PARAMETERS

Certains membres de l'équipe de recherche aimeraient adopter une approche plus analytique dans la définition du vocabulaire pivot [cf. section III.2] et ce groupe de paramètres peut être considéré comme un pas dans cette direction. Mais l'idée de ce type de paramètre, que Melchuk a étudiée en détail [MELCHUK (1967)] n'a été jusqu'ici adoptée chez nous qu'avec des restrictions. Melchuk fait l'hypothèse que ces paramètres sont des universaux: nous ne les introduisons que lorsqu'ils sont justifiés ouvertement par des synonymes ou des traductions. Ainsi dans les exemples suivants:

("causatif")	(i)	Eng. inform -- Fr. faire savoir
	(ii)	Eng. inform -- Eng. let know;
("inchoatif")	(iii)	Eng. go to sleep -- Fr. s'endormir,

les éléments soulignés sont une justification acceptable de la présence du paramètre en question.

3. STYLE et ATTITUDE

Il s'agit respectivement de "niveau de style" et de "jugement porté par le locuteur sur l'information cognitive qu'il transmet". Nous pensons que ces deux éléments font partie de la "signification" totale d'un lexème puisqu'elles sont reflétées dans le choix du locuteur parmi des lexèmes comportant la même valeur cognitive. Nous sommes conduits à cette extension de "signification" dès que nous voulons de bonnes traductions. Nous reconnaissons par exemple la différence entre "in future" (familier) et "henceforth" (rhétorique); ou entre "leave one's country" (attitude neutre) et "abandon one's country" (attitude de condamnation).

On peut se demander s'il existe des synonymes complets. Ainsi, lorsque l'on groupe des synonymes pour la traduction il est important de pouvoir décrire d'une part la "partie du sens" qui crée synonymie, et d'autre part la "partie de sens" qui distingue des synonymes partiels. La synonymie, pensons-nous, est fondée sur le contenu cognitif des lexèmes, tandis que STYLE, ATTITUDE et THESAURUS CATEGORY (voir ci-dessous) sont des paramètres de différentiation qui peuvent rendre une synonymie partielle sans la détruire entièrement.

4. THESAURUS CATEGORY

Il nous faut situer chaque usage d'un lexème dans un thesaurus structuré. Comme d'autres avant nous en T.A., nous nous sommes adressés à Roget, ou du moins à une collation moderne de son magnum opus [MAWSON (1946)]. Sa hiérarchie date du 19^{e} siècle, et au moins en partie est extra-linguistique et reflète une époque et une culture données. Elle est toutefois utile pour une description partielle des contextes déterminants pour les polysèmes, et même pour la reconnaissance des éléments principaux ("actants") dans notre grammaire de dépendance. La distinction théorique justifiée entre "facteurs linguistiques" et "facteurs culturels" dans la production du langage est trop vague pour être marquée en pratique.

Malheureusement, il n'y a pas de correspondant français au Roget, et nous n'avons pas les ressources et le temps pour compiler un nouveau thesaurus.

5. SYNTACTIC CATEGORY et REGIME

Ces détails n'apparaissent pas en ligne 4, c'est-à-dire qu'ils ne sont pas "persistants" dans le pivot. Les lexèmes du pivot ne comportent pas d'indications syntaxiques, car un concept exprimé en anglais par un verbe, par exemple, peut se trouver exprimé par un nom en français, etc. Mais d'autre part, les réalisations lexicales en anglais ou français doivent être classées en parties du discours pour que la syntaxe puisse être traitée. De plus, il nous faut attacher à chaque lexème la description des structures qu'il peut gouverner. Nous proposons le terme REGIME pour dénoter ces structures. Jusqu'ici, nous n'avons élaboré la description adéquate des régimes que pour les verbes et leurs "actants" (selon Tesnière et le CETA, avec des modifications). Mais la partie lexicographique permet l'élaboration future de la description des régimes pour les autres parties du discours.

Nous avons trouvé que les règles syntaxiques peuvent souvent s'appliquer à des catégories de thesaurus d'un niveau élevé dans la hiérarchie de Roget. Par exemple, il y a des formes de phrases caractéristiques des verbes de "communication humaine". Ces trouvailles sont importantes pour l'économie des grammaires, puisqu'elles tendent à confirmer l'hypothèse que la syntaxe n'est pas indépendante de la sémantique.

La description d'un régime a un but double: en plus de son utilité dans l'analyse de la phrase, elle constitue un contexte généralisé spécifiant un certain usage d'un lexème. Nous nous efforçons donc de généraliser de la sorte tous les contextes particuliers - c'est-à-dire les citations - que notre dictionnaire de base [HARRAPS (1967)], nous offre. Nous n'avons pas assez de place ici pour étudier le détail des sous-paramètres commandés par REGIME.

En conclusion, remarquons qu'il est très bien d'écrire une grammaire explicitant les relations entre les concepts d'une phrase ou d'un texte, mais que la moitié seulement du travail de T.A. est faite tant que les concepts eux-mêmes ne sont pas formellement définis.

VI - GRAMMAIRE D'ANALYSE DE L'ANGLAIS

La phase de consultation lexicale fournit à l'entrée de la grammaire d'analyse une séquence de mots du langage pivot assortis chacun d'une ou plusieurs catégories grammaticales. La sortie de la grammaire d'analyse est une séquence - munie d'un ordre hiérarchique - de mots du langage pivot qui représente les dépendances sémantiques de la phrase d'origine. Il y a une correspondance biunivoque entre les configurations sémantiques et les chaînes du langage pivot. Par conséquent les paraphrases seront représentées par une même chaîne du langage pivot, et les phrases n-ambiguïtés auront n représentations dans le langage pivot.

Les règles de la grammaire sont de deux types fondamentaux:

1. Règles qui assignent une catégorie grammaticale supérieure à une séquence de catégories grammaticales.

2. Règles qui permettent, effacent ou ajoutent des éléments dans une séquence.

Dans le formalisme W ces deux types de règles apparaissent dans deux parties séparées de la grammaire.

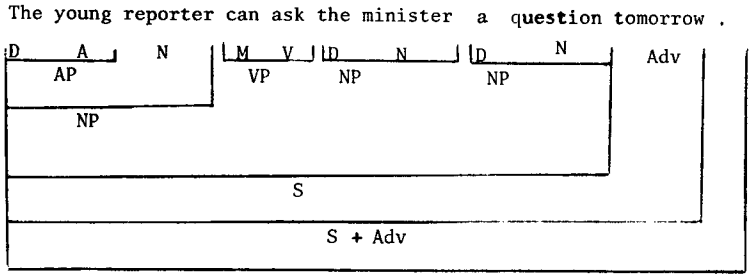
Les règles du type 1 effectuent une analyse en constituants immédiats de la chaîne d'entrée (ou de réarrangements de celle-ci effectués par des règles du type 2.) Pour qu'une chaîne pivot résulte de ce traitement, il est nécessaire que la catégorie "phrase complète" soit attribuée à la chaîne d'entrée dans son ensemble. Les règles de type 1

"essaient" d'assigner une catégorie à chaque sous chaîne de la chaîne d'entrée, mais seules seront conservées les attributions de catégories conduisant à l'attribution de la catégorie "phrase complète" à la chaîne d'entrée.

Les règles de type 2 ont deux fonctions. La première est de normaliser la représentation de variantes transformationnelles, qui seront ainsi représentées par la même chaîne pivot. La seconde est de réarranger les parties d'une chaîne selon un ordre hiérarchique et d'introduire des éléments de parenthétisation et d'étiquetage.

La grammaire actuelle a des règles spécifiant les structures possibles des composants syntaxiques principaux: groupes nominaux et verbaux, phrase, adjonctions. Il y a aussi des règles spécifiant des restrictions d'ordre sémantiques entre les composants principaux. Pour le moment celles-ci sont assurées par un système d'accords entre le verbe et ses actants.

Les structures de constituants décrites par notre grammaire sont en constante expansion. Certaines des structures principales décrites sont les suivantes: adjectifs, déterminants, propositions relatives, modifieurs prépositionnels, adverbes (d'un seul mot), modaux, auxiliaires, circonstanciels, etc. Les règles de type 1 donnent en gros l'analyse suivante:



Phrase complète

Dans l'analyse les règles de type 1 s'appliquent non seulement à la chaîne d'entrée mais aussi au résultat de l'application de règles du type 2. Le diagramme précédent n'est donc qu'une approximation du traitement réel. [Pour un exemple d'analyse montrant les règles utilisées à chaque étape voir RAPPORT SEMESTRIEL (1969)].

Les accords quasi sémantiques spécifiant les restrictions de cooccurrence entre un verbe et ses actants nous permettent d'éliminer certains cas d'ambiguïté syntaxique. Par exemple nous pouvons ainsi choisir des structures différentes pour:

The witness to the accident that occurred at the corner

et

The witness to the accident that spoke to the reporters

Les règles du type 2 qui éliminent les variantes paraphrastiques le font par réécriture des variantes en une même chaîne. Ainsi, les variantes:

The man to whom I gave it

The man whom I gave it to

The man who I gave it to

The man that I gave it to

sont toutes réécrites dans la même forme canonique, et traitées à partir de ce moment de la même façon. De cette manière, les variantes syntaxiques sans signification sémantique ne sont pas conservées jusqu'à la chaîne en langage pivot.

VII - SYNTAXE FRANCAISE

Le but de la génération du français est d'obtenir une expression adéquate des structures sémantiques codées dans les chaînes du langage pivot, qui soit aussi proche que possible du français (technique) standard. Il est évident que les raffinements stylistiques ne sont pas encore - et ne seront pas pour longtemps - à l'ordre du jour.

Nous recherchons essentiellement qu'une forme correcte de l'expression, on peut diviser la tâche en trois parties. Le français comporte des marques d'accord obligatoires et certaines contraintes d'ordre des éléments. De plus, il faut générer les formes correctes des lexèmes pourvus de leurs marques grammaticales.

Nous avons dû pour cela diviser la génération du français en quatre phases.

La première (I) détache de la structure sémantique codée dans les chaînes - pivot les éléments abstraits codant les lexèmes, et les remplace par des lexèmes français accompagnés de leur marqueurs grammaticaux inhérents (genre du nom, classes sémantiques et prépositions régies par le verbe, etc.)

La seconde (II) effectue une recomposition de la structure sémantique et une copie des marqueurs introduits en phase I en toutes les positions où ils sont requis par les règles d'accord du français.

La troisième (III) donne aux éléments l'ordre de surface du français. L'importance de cette phase est réduite dans une grande mesure par la décision de négliger pour l'instant toutes sortes de détails secondaires. C'est sur elle que nos efforts futurs devront porter si nous voulons améliorer la fidélité "stylistique" de la traduction.

Nous prévoyons la nécessité d'une quatrième phase séparée de la troisième. Cette phase IV serait proprement appelée "morphologie". Elle correspond grossièrement à la partie phonologique d'une grammaire générative, et n'est pour l'instant représentée que par quelques règles placées "en appendice" à la phase III. Des travaux effectués il y a quelques années par A. Dugas serviront de base à un traitement relativement simple de la morphologie.

VIII - SYSTEME-Q

A la lumière des expériences passées, il est apparu que le modèle mathématique de traduction que nous utilisions (grammaires-W) présentait certaines lacunes:

1. Difficulté de fractionner la phase d'analyse ou de génération en plusieurs phases.

2. Manque de souplesse pour manipuler certaines informations structurées sous forme d'arbre, en particulier lors de la phase de génération.

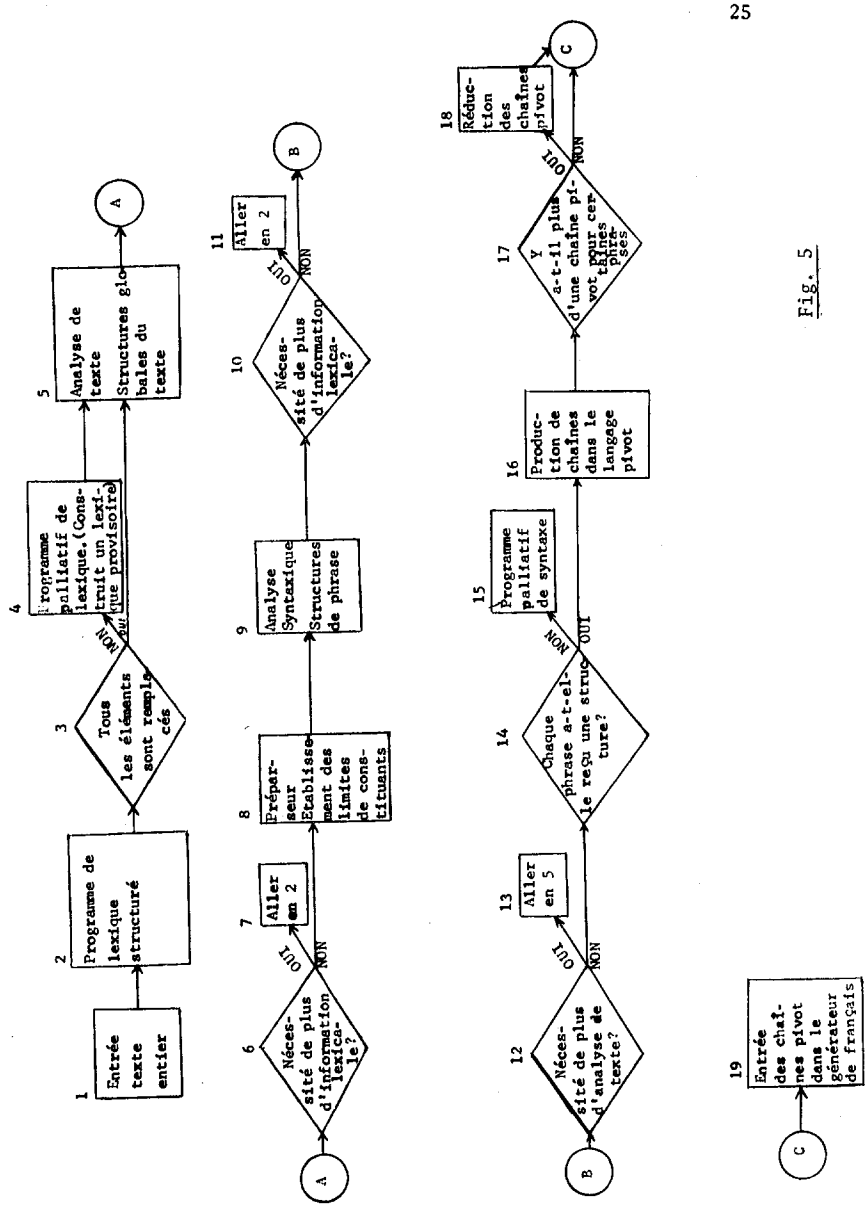

1
Entrée texte entier
2
Programme de lexique structuré
3
Tous les éléments sont remplacés
OUI
NON
4
Programme palliatif de lexique. (Construit un lexique provisoire)
5
Analyse de texte
Structures globales du texte
A
6
Nécessité de plus d'information lexicale?
OUI
NON
7
Aller en 2
8
Préparateur
Etablissement des limites de constituants
9
Analyse Syntaxique
Structures de phrase
10
Nécessité d'information lexicale?
OUI
NON
11
Aller en 2
B
12
Nécessité de plus d'analyse de texte?
OUI
NON
13
Aller en 5
14
Chaque phrase a-t-elle reçu une structure?
OUI
NON
15
Programme palliatif de syntaxe
16
Production de chaînes dans le langage pivot
17
Y a-t-il plus d'une chaîne pivot pour certaines phrases
OUI
NON
18
Réduction des chaînes pivot
C
19
Entrée des chaînes pivot dans le générateur de français

Fig. 5

3. Performances limitées quant à la taille des grammaires acceptées et du temps d'exécution des programmes. Ce dernier défaut découle d'ailleurs directement des défauts 1. et 2.

A. Colmerauer a donc commencé l'étude d'un nouveau type de grammaire plus adapté au but que nous nous proposons d'atteindre. Ces grammaires (systèmes-Q) seront essentiellement constituées de règles de réécriture générales pouvant non seulement s'appliquer à des chaînes mais aussi à des arbres. Un programme est en cours d'élaboration, qui permettra étant donné un texte ou une information structurée sous forme d'arbre, de lui appliquer un certain nombre de transformations décrites par une grammaire et d'obtenir un nouveau texte ou une nouvelle information structurée. En utilisant plusieurs fois ce même programme avec des grammaires différentes, on pourra alors enchaîner plusieurs phases d'analyse de l'anglais et plusieurs phases de génération du français. Il faut remarquer que, contrairement aux grammaires W, ce sera le même programme qui sera utilisé pour l'analyse et la génération. Ceci donnera plus de possibilités aux linguistes écrivant les grammaires: en effet, nous nous sommes aperçus que durant la phase d'analyse il était parfois nécessaire d'utiliser certains processus propres à la phase de génération et inversement durant la phase de génération, de réanalyser certaines parties afin de vérifier la grammaticalité du français généré.

Nous avons déjà commencé l'écriture d'une partie de ce nouveau système, qui est opérationnelle depuis juin 69. A cette date, nous avons donc pu commencer l'étude de la traduction automatique à une plus vaste échelle. Les grammaires-W qui sont déjà écrites seront très facilement réutilisables, le nouveau formalisme étant surtout une extension de ce que nous avons fait jusqu'à maintenant.

IX - ESQUISSE D'UN SYSTEME DE TRADUCTION

Les procédures de traduction décrites dans les sections précédentes comportent un certain nombre de limitations. Celles-ci nous ont conduits à examiner quelles seraient les extensions nécessaires de notre système. La figure 5 indique les types de traitement dont nous prévoyons la nécessité, ainsi que l'organisation du traitement. Faute de place, nous ne ferons pas de commentaires sur les parties qui sont déjà en cours de développement et n'ont qu'à être adaptées au système, par exemple 9: Programme d'analyse syntaxique. Nous ne parlerons donc que des sections qui en sont encore au stade théorique, mais pour lesquelles nous entrevoyons une réalisation possible, par exemple 5: Analyse de texte.

1. Entrée du texte entier

Le traitement actuel se fait phrase par phrase; le système proposé tiendrait compte de ce que l'analyse sémantique n'est possible qu'à partir d'un texte entier et non de phrases isolées.

2. Recherche dans un lexique sémantiquement structuré selon le principe de correspondance maximale

Le lexique proposé assignera aux chaînes d'entrée des ensembles de marqueurs sémantiques et syntaxiques (ces chaînes ne seront pas limitées à des mots isolés). La sortie du lexique, pour une chaîne donnée comporterait l'ensemble de ces marqueurs. Les sorties du lexique actuel sont constituées de mots du langage pivot assortis de tels marqueurs [cf. section V]. Au niveau de la phrase, un lexique sémantiquement structuré nous permettrait d'éliminer les sens de la phrase satisfaisant aux règles de la syntaxe, mais non à celles de la sémantique. Au niveau des relations entre phrases, il nous permettrait d'établir des relations entre les domaines sémantiques de diverses parties du texte.

3. et 4., 14. et 15. Programmes palliatifs

Ces programmes permettent de produire un résultat dans les cas où le système ne pourrait pas traiter un élément lexical ou syntaxique.

5. Programme d'analyse de texte

Ce programme comportera des règles concernant les relations entre les phrases d'un texte. Celles-ci assureront la cohérence du texte dans son ensemble. Parmi les tâches spécifiques que ce programme pourrait assurer, citons: 1. Clarification des références pronominales. 2. Désambiguation des éléments d'une phrase d'après d'autres éléments du texte. 3. Restauration de portions élidées du texte (par exemple restitution de l'agent effacé d'un passif).
On peut prévoir deux niveaux de règles différents:

1. Celles qui agissent sur des traits sémantiques pour éliminer des ambiguïtés ou établir des relations d'inclusion .

2. Celles qui traitent des interdépendances syntaxiques entre les phrases d'un même texte.

6., 10. et 11. Options de recyclage

On peut s'attendre que la complexité des relations entre choix lexical, structures de phrase et structures de texte requière parfois un traitement cyclique.

17. et 18. Réduction des chaînes pivot

Ce programme choisirait entre divers sens possibles qui n'auraient pas été éliminés par les programmes précédents, pour éviter la production de multiples traductions.

* - * - * - * - * - * - * - * - * - * - * - * - * - * - * - * - * - * - *

REFERENCES BIBLIOGRAPHIQUES

ANDREYEV (1965) : Andreyev, N. 'The intermediary language as the focal point of machine translation', in A.D. Booth (ed.), Machine Translation, Amsterdam: North Holland Publishing, pp. 1-27.

DE CHASTELLIER et COLMERAUER (1969) : G. de Chastellier et A. Colmerauer, 'W-Grammar', paper to be read at the 1969 ACM National Conference and Exposition, Aug. 26-28, San Francisco.

GOUGENHEIM (1954) : Gougenheim, G., et al., Le Français fondamental, 1er et 2e degrés, Paris: S.E.V.P.E.N.

HARRAPS (1967) : Mansion, J.E., Harraps Shorter French and English Dictionary, revised edn. by Ferlin and Forbes edited by D.M. and R.P.L. Ledésert, London: Harrap.

HARRIS (1968) : Harris, B., 'Entités lexicales abstraites de traduction', Recherche sur la traduction automatique, rapport trimestriel, x, 31 avril 1968, 16-34.

HOFMANN (1968) : Hofmann, T.R., Affixation-a new direction in transformational theory, Washington: ERIC/PEGS, July 18, PEGS Paper No. 53.

LEWIS (1961) : Lewis, N. (ed.), The New Pocket Roget's Thesaurus in Dictionary Form, New York: Washington Square.

MAWSON (1946) : Mawson, C.O.S., et Whiting, K.A. (eds.) Roget's Pocket Thesaurus, New York: Pocket Books.

MELCHUK (1967) : Mel'cuk, I.A., 'Semantičeskie parametry .....', in To Honor Roman Jakobson: Essays on the Occasion of His Seventieth Birthday, Hague: Mouton, pp. 1340-1361.

RAPPORT SEMESTRIEL (1969) : Projet de Traduction Automatique, Rapport semestriel, 12, Montréal: Université de Montréal et Conseil National de Recherches du Canada, 31 avril.

SPARCK JONES (1965) : Sparck Jones, K., 'Experiments in semantic classification', Mechanical Translation, viii, 3 and 4, June and Oct., 1965.