[*Automatic Translation of Languages*: NATO Summer School…July 1962]

UN SYSTÈME MORPHOLOGIQUE, COMPROMIS ENTRE LES FACILITÉS DE LA COMPILATION, LES RECHERCHES SYNTAXIQUES ET L'ADAPTATION À DE FUTURS PROGRAMMES DE T.A.*

L. DUPUIS

Centre d'Etudes pour la Traduction Automatique, Section de Paris, CETAP/CNRS

Résumé—Examen de certains problèmes généraux relatifs à l'organisation d'un dictionnaire pour la traduction automatique. Définition formelle élémentaire. Recherche d'une forme dans un dictionnaire. Structure des informations linguistiques enregistrées dans le dictionnaire, structures morphologiques et sémantiques. Principes d'un système d'analyse morphologique d'une langue flexionnelle. Applications à la langue russe.

1.

Etant donné l'importance et les difficultés des problèmes syntaxiques et sémantiques dans la T.A., le titre de cet exposé risque de paraître trivial. Il est bien connu, depuis les premières recherches faites dans les années 1954, que le traitement automatique de la morphologie d'une langue naturelle ne présente pas de difficultés insurmontables. Il suffit de bien vouloir y consacrer un peu de connaissances de la langue considérée et un peu d'ingéniosité dans la réalisation des programmes particuliers à ce genre de problèmes pour obtenir des résultats concrets et relativement satisfaisants.

En fait, notre intention est de situer l'objet de cet exposé dans le cadre plus général des études faites en vue de la réalisation d'un dictionnaire à l'usage de la T.A. et d'envisager, à ce propos, certains des problèmes relatifs à la compilation des informations linguistiques intéressant la T.A.

Les recherches faites ces dernières années à propos de la construction des grammaires montrent que la qualité des résultats obtenus par un programme d'analyse syntaxique utilisant une grammaire d'un type donné dépend étroitement de la quantité d'informations linguistiques qui est incorporée dans cette grammaire. Une grammaire faible associée à un programme de reconnaissance des structures apporte des résultats qui ne sont pas dénués d'intérêt pour la recherche linguistique, mais il est très malaisé d'enrichir et d'améliorer une telle grammaire par la suite (cf. [8]). Mieux vaut, semble-t-il, énoncer très tôt les contraintes sémantiques susceptibles d'éliminer les nombreuses structures manifestement fausses données par une grammaire faible, tout en reconnaissant que l'énoncé de ces contraintes est l'un des problèmes les plus difficiles de la T.A. En raison de l'arbitraire et de la motivation relative du signe linguistique, il est probable que ce problème n'admet pas de solution générale.

Ces résultats permettent de mieux situer la place du dictionnaire dans la T.A. Les performances techniques: compacité des informations enregistrées, vitesse de consultation, ne sont pas des facteurs essentiels, tout au moins dans un avenir immédiat; ce qui ne

* Presented at the NATO Advanced Study Institute on Automatic Translation of Languages: Venice, 15-31 July, 1962.

signifie pas qu'il s'agit là de problèmes faciles et que tout perfectionnement dans ce sens ne soit pas le bienvenu.

Par contre, la facilité de la compilation des informations linguistiques, la commodité de la consultation de ces informations, la mise en évidence et l'utilisation de structures relatives à ces informations, sont des facteurs susceptibles de remédier à la difficulté théorique des problèmes sémantiques, en particulier la construction de catégories sémantiques et au fait indiqué ci-dessus à propos des programmes de reconnaissance des structures. Le dictionnaire n'est pas un simple répertoire que l'on consulte pour des raisons de commodité; il fait partie intégrante des programmes de traduction et il doit faciliter au maximum l'apport des nombreuses informations linguistiques nécessaires à la qualité des traductions, automatiques ou non.

2.
DÉFINITION FORMELLE ÉLÉMENTAIRE D'UN DICTIONNAIRE

Un dictionnaire est une suite finie de couples:

$$(F_1S_1)\ldots(F_i.S_i)\ldots(F_nS_n).$$

Chaque F_i est une *forme graphique* ou assemblage linéaire de symboles élémentaires appartenant à un alphabet donné. Cet assemblage de symboles est construit suivant des règles formelles déterminées.

Chaque S_i est un ensemble d'informations associées à la forme F_i.

Consulter le dictionnaire pour une forme F_i, c'est rechercher la forme F_i dans la suite $F_1 S_1 , . . F_n S_n$, et lire l'ensemble des informations associé à la forme F_i.

Pour un langage naturel, le couple *(F, S)* est un mot (ou forme linguistique) F est le *signifiant* de ce mot (les symboles élémentaires sont les lettres de l'alphabet de ce langage), S est le *signifié* de ce mot (représenté pratiquement par l'ensemble des informations morphologiques, syntaxiques et sémantiques associées à ce mot et par une définition dans le cas d'un dictionnaire monolingue ou par une liste d'équivalents de la langue-cible dans le cas d'un dictionnaire bilingue).

Constituter un dictionnaire automatique, c'est enregistrer la suite $F_1S_1, . . . F_n S_n$, dans la mémoire d'un calculateur électronique et établir un programme-machine (programme de consultation) pour effectuer l'opération de recherche d'une forme F_i et l'opération de lecture des informations S_i.

Un article d'un tel dictionnaire est le couple *(F,* S), la forme graphique est son *en-tête,* l'ensemble des informations S est son *contenu,*

Le temps de consultation T est l'intervalle de temps *moyen* pour trouver dans le dictionnaire une forme quelconque F_i d'un texte écrit et pour lire l'ensemble des informations correspondantes S_i (qui peut être nul, si F_i n'est pas dans le dictionnaire). Les performances relatives à la vitesse de consultation ne sont significatives que par rapport au volume V des informations linguistiques enregistrées dans le dictionnaire et représentées par la suite $F_1S_1\ldots F_n S_n$.

PROBLEMES IMMÉDIATS RELATIFS A LA DÉFINITION PRÉCÉDENTE

En principe, la forme F est une suite de lettres de l'alphabet de la langue considérée, comprise entre deux intervalles ou blancs. S est l'ensemble de toutes les informations linguistiques que l'on peut énoncer à propos de cette forme F.

Il est bien connu que cette définition demande certaines précautions :

(1) dans l'énumération des lettres de l'alphabet, on doit tenir compte de certains signes diacritiques, par exemple, en français, les accents, la cédille, le tréma (ainsi, l'alphabet français a 39 lettres et non 26); certains signes de ponctuation sont susceptibles d'être considérés comme des éléments de la forme *F* (par exemple, le trait-d'union, l'apostrophe).

(2) les mots homographes sont des mots différents qui ont une forme graphique *F* commune mais des signifiés S_1 , S_2... différents. Les nécessités de l'automatisation imposent en général que l'on énonce simultanément ces divers signifiés à propos de la forme *F,* mais il est important de séparer nettement les informations linguistiques correspondantes, même si elles sont très voisines.

(3) l'unité de signification qui correspond naturellement au symbole formel *S* précédent, n'est pas nécessairement associée à une forme graphique *F* limitée par deux intervalles. Elle peut être associée, soit à une forme plus petite: mots composés formés de deux unités de signification ou plus, mots dérivés obtenus par l'adjonction de préfixes ou suffixes porteurs d'une certaine signification, mots fléchis par l'adjonction d'une désinence, soit à une forme plus grande. C'est le cas, non seulement des idiotismes mais aussi de nombreux groupes de mots techniques qui sans être de véritables idiotismes, constituent de véritables unités de signification dans lesquels il n'y a pas lieu de distinguer a priori des mots composants, sinon dans un lexique bien délimité et pour des raisons de simplification.

En fait, toutes ces particularités proviennent du même fait linguistique qu'on retrouve constamment: il n'est pas possible de séparer arbitrairement le signifiant *F* et le signifié *S*, par exemple, définir au préalable des formes admissibles *F,* puis les informations correspondantes *S.* Si on le fait, c'est uniquement pour des raisons de commodité, mais en définitive une définition formelle de *F* n'est valable qu'à une étape de la recherche et pour un mode d'exploitation déterminé.

RECHERCHE D'UNE FORME *F* DANS LE DICTIONNAIRE

Chaque couple *(F, S)* est enregistré dans une certaine zone de mémoire, formée d'un ou plusieurs mots-machine et repérée par une adresse A (par exemple, l'adresse du premier mot-machine de la zone) et des indications de format des enregistrements. Il s'agit, pour une forme *F* donnée, de retrouver l'adresse *A* correspondante.

De nombreux procédés ont été proposés pour effectuer cette recherche :

comparaison mot à mot. On compare la forme *F* successivement à chacune des formes du dictionnaire $F_1 \ldots F_N$ jusqu'au moment où on trouve la forme recherchée *F.*

Ce procédé est le plus simple, mais il n'est pas rapide. Le nombre des comparaisons nécessaires pour retrouver une forme déterminée est égal en moyenne à la moitié du nombre des formes enregistrées. Diverses variantes ont été proposées pour accélérer le temps de consultation moyen :

(1) Si le dictionnaire est placé sur une bande magnétique, les formes sont rangées dans l'ordre alphabétique et avant consultation, les formes du texte sont triées dans le même ordre. Après la consultation (faite en général par sections alphabétiques et par sections suivant la longueur des mots), ces formes sont retriées dans l'ordre du texte. Le temps de consultation est alors limité par le temps nécessaire pour effectuer ces tris et aussi par le nombre de formes du texte que l'on peut trier en une seule fois. Pour éviter les tris,

on peut aussi ranger les formes du dictionnaire par ordre de fréquences décroissantes, mais dans ce cas, il faut consulter tout le dictionnaire pour s'assurer qu'une forme ne s'y trouve pas.

(2) Le dictionnaire est rangé sur les pistes d'un disque d'une mémoire photoscopique (machine de G. King). La recherche d'une forme est faite par approximations successives. Ce procédé permet une application simple du principe du 'longest matching' pour la recherche des expressions idiomatiques comme 'parce que'.

calcul d'adresse. En machine, chaque lettre est codée par un nombre binaire ou décimal. La forme F est ainsi représentée par un nombre élevé (par exemple, un nombre de 48 chiffres décimaux, pour un mot de 24 lettres et un code décimal). On calcule l'adresse de la forme, à partir du nombre précédent, par un algorithme déterminé.

Divers algorithmes ont été envisagés pour ce calcul d'adresse. Ils sont basés sur la redondance d'information contenue dans les formes des langages naturels, compte tenu de la dépendance des symboles successifs de ces formes et sur divers renseignements statistiques relatifs à la longueur des mots, les fréquences des premières et dernières lettres (cf. [6]). Aucun d'eux n'est rigoureux. Le mieux que l'on puisse en attendre est de déterminer une zone de petites dimensions dans laquelle se trouve la forme cherchée, la recherche définitive étant faite par comparaison.

Une variante de ce procédé est celle utilisée par le CETAP pour l'adressage des mots d'un dictionnaire placés sur une unité de disques magnétiques 355. Pour une forme donnée F, on calcule à l'aide d'un algorithme très simple, 2 nombres pseudo-aléatoires A_1 et A_2. Si une seule forme F donne lieu au résultat A_1 cette forme est rangée sur la piste d'adresse A_1. Si plusieurs formes $F_1\ F_2\ F_3\ \ldots$ donnent lieu au résultat A_1; la piste d'adresse A_1 contient une table $T(A_2)$ dont les arguments sont les nombres A_2 et les fonctions, les adresses des pistes où sont rangées effectivement les formes $F_1\ F_2\ F_3$.... Ce procédé n'est pas le plus économique du point de vue temps de consultation et compacité des informations mais il est très simple à mettre en œuvre et convient très bien pour un dictionnaire pas trop volumineux dont les articles sont susceptibles d'être modifiés dans le temps, comme c'est le cas au stade de la recherche.

Une autre variante utilisée avec un dictionnaire enregistré sur une bande magnétique consiste à placer les formes du texte en mémoire rapide à des adresses calculées par un algorithme déterminé et à utiliser le même algorithme pour rechercher si chacune des formes du dictionnaire coincide avec une forme du texte (cf. [2]).

cheminement dans un arbre ou comparaison lettre par lettre. On découpe le mot lettre par lettre. Une table indique la zone de la mémoire où se trouvent rangées les formes dont la première lettre est une lettre déterminée; une deuxième table indique la zone de la mémoire où se trouvent rangées les formes débutant par 2 lettres déterminées et ainsi de suite. Ce procédé permet la mise en œuvre des programmes de liste. Il se rapproche du précédent par les faits suivants :

La construction des tables est simplifiée par la dépendance qui existe entre les symboles successifs d'une forme.

Certains doublets ou triplets n'existent pas dans la langue—une comparaison lettre par lettre de la forme entière n'est pas économique. On utilise le procédé seulement pour les 4 ou 5 premières lettres, ce qui constitue un prérepérage de la zone où se trouve la forme cherchée. La recherche définitive de la forme est faite par comparaison (cf. [3]).

conclusion. Ces diverses méthodes ont leurs avantages et inconvénients. En général, les réalisations effectives résultent de dosages variables de ces méthodes. Elles dépendent

du type de machine utilisé et des procédés utilisés pour comprimer le volume réel des informations enregistrées (par exemple, dictionnaire de bases, voir ci-après).

Une méthode utilisée consiste à découper le dictionnaire en tranches verticales: morphologie, syntaxe, sémantique. Le dictionnaire morphologique ne contient que des formes et peut être considéré comme un procédé d'adressage du type précédent. Dans cette méthode, on doit tenir compte soigneusement du fait que les informations morphologiques, syntaxiques et sémantiques relatives à une forme sont susceptibles d'être de longueurs très différentes et par suite il n'y a pas de correspondance simple entre les adresses des enregistrements relatifs à ces diverses informations.

3.
STRUCTURE DES INFORMATIONS LINGUISTIQUES ENREGISTRÉES DANS LE DICTIONNAIRE

La structure élémentaire d'un dictionnaire est celle indiquée ci-dessus: $(F_1\ S_1)\ \ldots\ (F_i,\ S_i)\ \ldots\ (F_n,\ S_n)$.

Elle est caractérisée par une correspondance biunivoque entre l'ensemble des formes (Fi) et l'ensemble des (S_i). (F, S) est un mot, F est le signifiant, S est le signifié. Cette organisation est celle d'un *dictionnaire de formes*. Pour la recherche, elle n'est pas dénuée d'intérêt. Elle permet d'enrichir progressivement et commodément le contenu S d'un article de dictionnaire au fur et à mesure des progrès de la recherche, sans que l'on soit astreint à modifier parallèlement le contenu des autres articles.

Mais on doit remarquer qu'une telle organisation est celle d'un répertoire d'informations, analogue à celle d'un annuaire téléphonique. Le lexique d'une langue naturelle n'est pas une simple nomenclature mais un système où tout se tient et où les mots ne prennent leur valeur qu'en fonction les uns des autres. Ce fait, bien connu des lexicographes, conduit à étudier le problème plus général de la structure des informations linguistiques contenues dans un dictionnaire. En particulier, on doit tenir compte des faits signalés ci-dessus à propos de la non correspondance entre les formes graphiques telles qu'on les définit habituellement et les unités de signification.

Cette structure présente deux aspects: structure morphologique et structure sémantique. *Une structure morphologique* est caractérisée par l'existence de relations entre les segments obtenus par découpage des diverses formes de la langue et corrélativement entre les informations linguistiques que l'on peut associer à ces segments. *Une structure sémantique* est caractérisée par l'existence de relations entre les signifiés de la langue et corrélativement entre les formes associées à ces segments.

STRUCTURES MORPHOLOGIQUES

Supposons que toutes les formes d'un ensemble de k mots de la langue admettent un élément commun B

$$F_1 S_1 = (B D_1) S_1$$

$$F_k S_k = (B D_k) S_k.$$

Cet ensemble de k mots constitue une *unité morphologique* relative à la base B et à la suite des affixes $D_1 \ldots D_k$. Cette unité morphologique peut constituer un seul article de dictionnaire si l'on prend comme en-tête de l'article, la base B et comme contenu la suite $D_1 S_1 \ldots D_k S_k$.

Une telle structure est avantageuse si l'ensemble des $S_1 \ldots S_k$ admet une décomposition

analogue: $S_i = W_i\ S$, S étant l'élément commun aux divers $S_1 \ldots S_k$ et $W_1 \ldots W_k$ des informations linguistiques particulières à chaque mot. Dans ce cas, l'article relatif à l'unité morphologique considérée peut s'écrire:

$$BS(D_1\ W1 \ldots D_k\ W_k)$$

la suite $\Pi = D_1\ W_1 \ldots D_k\ W_k$ est *un paradigme linguistique* si elle est commune à plusieurs unités morphologiques. Chacun des articles de dictionnaires relatifs aux mots de ces unités morphologiques peut s'écrire: $B\ S\ \Pi$ où Π est un simple numéro d'ordre relatif au paradigme linguistique. Ce numéro d'ordre et le contenu correspondant c'est-à-dire la suite $D_1\ W_1 \ldots D_k\ W_k$ constituent un seul article qu'il suffit d'écrire une seule fois, quelque soit le nombre d'unités morphologiques auxquelles il se rapporte.

Une telle structure est très générale. Elle dérive naturellement de la notion habituelle de paradigme, utilisée pour les déclinaisons et les conjugaisons. Dans ce cas les affixes D_i sont les désinences habituelles, les W_i sont les modalités associées à ces désinences : cas, nombre personne….

Mais une telle structure est valable également pour des affixes autres que les désinences habituelles: préfixes, infixes, suffixes. Par exemple, dans la langue russe, certains verbes admettent 4 participes dont les bases dérivent respectivement de la base non participe par l'adjonction des suffixes: ющ (participe présent actif), *В*ш (participe passé actif), *EM* (participe présent passif), *HH* (participe passé passif). Si ϕ désigne l'absence de suffixe, on a le paradigme suivant relatif aux bases des participes :

ϕ (non part.) ющ (P. pr a), *В*ш (p. pa a), *EM* (P. pr p) *HH* (p. pa p).

Une catégorie morphologique est l'ensemble de toutes les unités morphologiques définies par des informations linguistiques W_i déterminées. Par exemple, en allemand ou en russe, la catégorie morphologique substantif est l'ensemble des unités morphologiques pour lesquelles la suite $P = W_1 \ldots W_k$ est définie par toutes les combinaisons possibles de cas et de nombre (en allemand $k = 4 \times 2$, en russe $k = 12 = 6 \times 2$).

En général, une catégorie morphologique coincide avec une catégorie grammaticale habituelle, mais ce n'est pas obligatoire. Par exemple, les formes adjectivales russes appartiennent à deux catégories morphologiques distinctes: adjectif forme longue et adjectif forme courte. Les adjectifs pronominaux et les nombres ordinaux appartiennent à la catégorie morphologique adjectif longue tandis que les nombres cardinaux constituent une catégorie à part.

CAS PARTICULIERS

(1) Les diverses formes $F_1 \ldots F_k$ d'une même unité morphologique n'ont pas d'éléments communs mais se répartissent en plusieurs groupes ayant chacun un élément commun: $B_1\ B_2 \ldots$ Dans ce cas, il faut considérer des *paradigmes partiels* relatifs à ces diverses bases. L'unité morphologique est ainsi constituée des sous-unités $B_1\ \Pi_1\ S_1\ B_2\ \Pi_2\ S_1 \ldots$

(2) Certaines formes correspondantes à la suite $W_1 \ldots W_n$ n'existent pas dans la langue. Dans ce cas, il faut considérer des *paradigmes défectifs* (par exemple, pour un substantif, des paradigmes défectifs au singulier ou au pluriel.

La mise en évidence de telles structures morphologiques permet de constituer et d'utiliser commodément un *dictionnaire de bases* dont l'intérêt pour les langues flexionnelles est bien connu: on évite d'enregistrer plusieurs fois des informations communes à divers mots d'une même unité morphologique: la base B et les informations linguistiques communes S. De

même, les informations relatives à un paradigme sont enregistrées une seule fois, quel que soit le nombre d'unités morphologiques auxquelles elles sont susceptibles de se rapporter.

Un autre intérêt, propre au système proposé, réside dans la notion de paradigme: si l'on prend certaines précautions (voir ci-après) un paradigme constitue une description complète et correcte des informations linguistiques particulières à chaque élément d'une unité morphologique qui admet ce paradigme (informations dites morphologiques). Par suite, le découpage des formes réalisé suivant ce modèle ne donne lieu ni à des pertes d'informations linguistiques ni à des associations incompatibles avec la langue. D'autre part, une telle description permet un indexage et une vérification commode et efficace des informations enregistrées dans le dictionnaire-machine. Elle facilite les échanges de dictionnaires avec d'autres équipes de T.A. n'utilisant pas nécessairement le même système, car il est possible de reconstituer aisément le dictionnaire de formes à partir duquel on peut faire dériver tout autre dictionnaire de bases.

Enfin, un tel système facilite l'interpolation des mots non trouvés dans le dictionnaire. L'ensemble des paradigmes recensés permet de constituer une table relativement complète des couples: désinences, informations grammaticales correspondantes. En présence d'une forme non trouvée, une consultation de cette table donne un minimum d'informations grammaticales pour cette forme, par simple examen des désinences admissibles.

La constitution et l'utilisation des listes de paradigmes demandent certaines précautions :

(1) *du point de vue linguistique,* on doit vérifier que pour toutes les unités morphologiques qui admettent un paradigme déterminé Π, la suite $B\ S\ \Pi = B\ S\ (D_1\ W_1 \ldots D_n\ W_n$ constitue une description suffisante de tous les éléments de l'unité morphologique correspondante. Par exemple :

la forme relative à un certain *W* d'un substantif particulier peut faire partie d'un idiotisme dont les propriétés sémantiques sont différentes des propriétés sémantiques habituelles du substantif correspondant.

les formes participiales d'un verbe n'admettent généralement pas les mêmes régimes que les formes non participes du même verbe.

Si de telles singularités linguistiques ne sont pas très fréquentes dans une catégorie morphologique déterminée, il suffit d'ajouter aux informations communes *S,* les informations linguistiques particulières à une forme, au prix d'une certaine complication du contenu de l'article et de sa consultation.

Sinon, il est préférable de renoncer provisoirement à l'utilisation de paradigmes linguistiques pour la catégorie morphologique en question (par exemple, dans le dictionnaire CETAP actuel, on a constitué pour un verbe autant d'articles qu'il y a de bases de participes).

(2) *du point de vue de la programmation.* L'en-tête de l'article relatif à l'unité morphologique $B\ S\ \Pi$ est la base *B.* Ceci suppose qu'avant consultation du dictionnaire pour une forme *F,* le programme a effectué le découpage $F = B\ D$. Or dans certains cas, plusieurs découpages sont possibles $F = B_1\ S_1\ B_2\ D_2, \ldots$ dont un seul est valable en général (sauf dans le cas où la forme *F* est équivalente à plusieurs formes homographes suivant qu'il s'agit de D_1 ou $D_2 \ldots$). Ce fait entraine une augmentation du temps de consultation moyen sensiblement proportionnelle au nombre des découpages déclanchant les recherches. Diverses méthodes ont été proposées pour ramener ce temps à une valeur voisine de celle relative à un seul découpage. La plupart d'entre elles reviennent à choisir convenablement la liste des désinences.

Par exemple, si la désinence D_1 est contenue à droite dans la désinence D_2 : $D_2 = dD_1$ (en français ons = on – s), une forme *F* terminée par une suite de lettres identiques à D_2

peut être enregistrée dans l'un des articles de dictionnaire suivants (selon le paradigme linguistique auquel elle appartient)

$$A_1 = B_1\ (D_1\ D_3) = (B_2 d)\ (D_1\ D_3)$$
$$A_2 = B_2\ (D_2\ ,\ D_4) = B_2\ (dD_1\ D4).$$

D_3 et D_4 étant des désinences quelconques des paradigmes respectivement relatifs aux bases B_1 et B_2.

Pour ne pas avoir de faux splittages, on peut:

(1) soit supprimer la désinence $D_2 = dD_1$ c'est-à-dire diminuer la longueur et le nombre des désinences admissibles. En général ceci revient à augmenter la longueur et le nombre des bases. Par exemple, si A_2 existe, il est nécessaire de le dédoubler puisqu'en général $D_4 - d$ ne sera pas une désinence admissible. A la limite on aurait un dictionnaire de formes.

(2) soit créer la désinence artificielle $D_5 = dD_3$ c'est-à-dire augmenter la longueur et le nombre des désinences et diminuer la longueur des bases.

Si A_1 existe, il doit être écrit suivant la forme:

$$A_1 = B_2\ (dD_1\ ,\ dD_3)$$

Pour éviter un faux splittage relatif à la désinence D_1 il faut admettre, que dans le cas où plusieurs découpages d'une forme sont possibles, c'est celui relatif à la désinence la plus longue qui est le seul valable. C'est le principe du splittage maximum.

Ce principe donne des résultats satisfaisants pour la langue russe. Au moyen de l'introduction de désinences artificielles convenables, on peut espérer réduire le nombre des cas de faux splittages, à une valeur faible, moins de 10% du nombre des formes différentes de la langue. Ce choix des désinences artificielles est fait commodément à l'aide du système de paradigmes décrit ci-dessus.

STRUCTURES SÉMANTIQUES

Une suite de k mots de la langue :

$$U = (F_1\ S_1)\ (F_2\ S_2)\ \ldots\ (F_k\ S_k)$$

constitue *une unité de signification* si la signification globale S que l'on doit associer à la suite formelle $F = F_1\ F_2\ \ldots\ F_k$ ne dérive pas des significations usuelles $S_1\ \ldots\ S_k$ associées respectivement à chacune des formes $F_1 \ldots$ et de l'application de règles grammaticales très générales, mais résulte au contraire d'une convention particulière adoptée par les utilisateurs de la suite F.

Habituellement, cette notion est réservée à la catégorie très particulière des idiotismes. En fait, elle est beaucoup plus générale. Par exemple, les significations des deux suites formelles : 'nombre rationnel', 'extraire une racine' ne dérivent pas des significations usuelles des mots isolés: 'nombre', 'rationnel', 'extraire', 'racine' et de la connaissance de leurs catégories grammaticales usuelles. Elles résultent plus précisément des définitions données par les mathématiciens.

Par suite, une telle unité de signification doit constituer un article de dictionnaire dont l'en-tête est la suite des formes $F = F_1\ \ldots\ F_k$ et le contenu, les informations linguistiques que l'on doit associer à la signification globale S. Dans un dictionnaire usuel, on trouve effectivement de tels articles, mais en général chacun d'eux est incorporé dans un article relatif à une forme isolée, par exemple F_1.

Si l'on s'astreint, pour des raisons de commodité, à constituer seulement des articles de dictionnaires simples du type $(F_i\ S_i)$, les signifiés de ces mots isolés sont les significations

$S_1 \ldots S_k$ que l'on doit associer respectivement aux formes $F_1 \ldots F_k$, compte tenu de la signification globale.

Par exemple, si l'on peut établir une correspondance entre ces formes et les formes $F'_1 \ldots F'_k$ de la traduction de l'unité de signification, les signifiés $S_1 \ldots S_k$ seront représentés par les équivalents respectifs $F'_1 \ldots F'_k$ dans la langue-cible des formes $F_1 \ldots F_k$ de la langue-source.

En général, une forme déterminée F_1 peut participer à plusieurs unités de signification et par suite admettre plusieurs signifiés $S_1\ S_2\ S_3 \ldots$ (polysémie). En raison de l'arbitraire du signe linguistique dans les langues naturelles et du petit nombre de formes effectivement utilisées dans une langue par rapport au nombre d'unités de signification l'existence de telles polysémies est un fait très général. Contrairement à une hypothèse souvent admise, un mot isolé est, a priori, *polysème,* s'il ne l'est pas, c'est un cas particulier dont il faut profiter mais qu'on ne peut ériger en règle générale.

La résolution de telles polysémies ne peut être envisagée que dans la mesure où l'on considère des relations entre les signifiés de la langue, c'est-à-dire, si l'on définit des *structures sémantiques* sur l'ensemble des mots de cette langue.

Considérons un ensemble de *n* unités de signification :

$$U_1 = (F^1{}_1 \ldots F^{k1}{}_1)\ S_1\ G_1$$
$$U_n = (F^1{}_n \ldots F^{kn}{}_n)\ S_n\ G_n$$

chaque unité U_i étant caractérisée par une suite formelle $F^1{}_i \ldots F^{ki}{}_i$, une signification globale S_i est une structure grammaticale (au sens habituel) G_i.

On définit *une structure sémantique élémentaire* sur cet ensemble si, compte tenu des diverse significations globales $S_1 \ldots S_n$, on peut associer à chacune des formes isolées $F^j{}_i$ un signifié unique $S^j{}_i$, valable dans toutes les unités de signification où $F^j{}_i$ apparaît.

Une telle structure définit une *catégorie sémantique* C_1 que l'on peut associer à chacun des mots isolés de l'ensemble considéré. Un mot polysème est un mot susceptible d'appartenir à plusieurs catégories sémantiques distinctes $C_1\ C_2 \ldots$

Une telle catégorie sémantique, définie par des relations entre les signifiés des mots isolés d'un ensemble d'unités de signification, est voisine de celle obtenue en considérant les signifiés des mots d'un domaine technique particulier (microglossaire). Elle peut être plus restreinte car certains mots d'un domaine technique sont susceptibles d'être polysèmes par exemple, le mot 'entier' n'a pas la même signification dans 'nombre entier' et 'fonction entière'. Elle peut être plus étendue car un mot technique peut conserver la même signification en dehors de son domaine de définition: le mot 'nombre' garde en général la même signification en dehors des textes mathématiques.

Le paradigme sémantique d'un mot *M* pour une relation grammaticale déterminée *G* est l'ensemble des mots appartenant à la même catégorie sémantique *C* que le mot *M* et susceptibles d'être associés au mot *M* par la relation grammaticale *G*.

Cette notion de paradigme sémantique précise les informations sémantiques que l'on peut trouver dans un dictionnaire habituel. Par exemple, dans l'expression:

'donner quelque chose à quelqu'un'

'quelque chose' désigne l'ensemble des objets qui peuvent être associés au mot 'donner' (muni de la signification: faire don de) par la relation grammaticale: verbe—complément direct d'objet. Cet ensemble est donc le paradigme sémantique du mot donner pour cette relation grammaticale. Cette notion permet de classer les unités de signification d'après la

longueur du paradigme sémantique de chacun de leurs mots composants, c'est à dire suivant le nombre des mots qui composent ce paradigme.

(1) *Idiotismes.* Le paradigme sémantique de chacun des mots composants est réduit à un seul mot (ou à un nombre très restrient de mots). C'est le cas d'expressions telles que 'parce que', 'avaler la pilule' (dans le sens se déterminer à une chose pénible) et de quelques expressions techniques.

(2) *Expressions techniques.* Le paradigme sémantique de chacun des mots composants peut comporter un nombre relativement élevé de mots, mais ces mots appartiennent à un domaine technique bien déterminé et sont susceptibles d'être énumérés: par exemple nombre entier, naturel, rationnel, irrationnel, complexe, imaginaire . . . ensemble borné, ouvert, fermé, dense, dénombrable....

(3) *Expressions du langage courant.* Ou expressions communes à l'ensemble des individus utilisant le langage considéré. Il s'agit d'expressions telles que: 'construire une maison', 'écrire une lettre', 'donner quelque chose à quelqu'un'. Dans de telles expressions, les mots composants ont une signification très générale et on ne peut prétendre énumérer tous les éléments de leur paradigme. Dans ce cas, un paradigme ne peut être construit que par extrapolation, à partir d'une suite d'éléments déjà recensés, en essayant de dégager les propriétés sémantiques générales de ses éléments (être animé, inanimé, action faite dans un but déterminé . . .).

4.
UN EXEMPLE D'APPLICATION: STRUCTURE MORPHOLOGIQUE RELATIVE AUX DÉCLINAISONS ET AUX CONJUGAISONS D'UNE LANGUE FLEXIONNELLE

Cette structure a pour objet de faire entrer dans le cadre général des structures définies ci-dessus, les déclinaisons et les conjugaisons étudiées dans les grammaires traditionnelles.

Elle est relative à des formes graphiques élémentaires : chacune des formes considérées est une suite de lettres de l'alphabet limitée par 2 intervalles ou blancs.

On suppose que le signifié associé à chacune de ces formes appartient à une catégorie grammaticale déterminée G (l'une des catégories traditionnelles, nom, adjectif, verbe . . .). Dans le cas où il est susceptible d'appartenir à plusieurs catégories (homographie externe), on suppose qu'il s'agit de plusieurs mots distincts.

Les unités morphologiques considérées sont des unités relatives aux modalités de la signification: cas, genre, nombre.… Une modalité est considérée comme une *variable grammaticale* V_i susceptible de prendre un nombre fini de valeurs v^k_i. La réunion des valeurs des variables grammaticales particulières à un mot sera notée : $W_i = v^i_i\, v^k_2\, v^j_3$....

Une unité morphologique sera l'ensemble des formes relatives à une suite $W_1 \ldots W_k$ définie comme suit:

Soit V un ensemble de variables grammaticales, par exemple: $V = V_1 V_2\ V_3\ V_4$ = genre, animation, cas, nombre (variables grammaticales d'un substantif russe). La suite $W_1 \ldots W_k$ est caractérisée comme suit: certaines variables conservent une valeur constante V_1 = masculin V_2 = inanimé. Les autres variables prennent toutes les valeurs possibles V_3 = nominatif, génitif, datif,... V_4 = singulier, pluriel.

Chaque W_i de la suite $W_1 \ldots W_k$ est déterminé par une combinaison particulière de ces valeurs :

W_1 = masc. inan. nom. sing.

W_{12} = masc. inan. locatif pluriel.

Les formes $F_1 \ldots F_k$ associées à cette suite de valeurs de variables grammaticales sont les diverses formes d'un mot linguistique déterminé (dans le sens habituel de ce terme), c'est-à-dire les formes obtenues dans le cours de la déclinaison ou de la conjugaison d'un mot.

En première approximation, cet ensemble de formes constitue bien une unité morphologique dans le sens défini ci-dessus:

(1) en principe, les informations linguistiques associées à ces formes (autres que celles associées aux *W)* sont toutes identiques.

(2) en général, ces formes admettent un élément commun qui est la base (ou radical) du mot considéré. Dans certains cas, l'unité morphologique peut admettre plusieurs bases (par exemple, en français, le mot: œil, yeux). Les éléments non communs sont les désinences habituelles $D_1 D_2 \ldots D_k$. Le paradigme correspondant est la suite $D_1 W_1 \ldots D_k W_k$.

Paradigme réduit dans certains cas, les formes $F_i F_j$ ... relatives à des W_i, W_j distincts sont toutes identiques *(homographie interne).* Dans ce cas, il est commode de considérer une forme unique $F_h = F_i = F_j$ et d'associer à cette forme F_h un W_h unique formé de la réunion de W_i et de W_j: $W_h = W_i W_j$. Un tel W_h est dit *ambigu.*

Un paradigme dans lequel toutes les formes sont différentes est dit *paradigme réduit.*

Un paradigme-machine est la suite formelle

$$\Pi = \delta_1 W_1, \delta_2 W_2, \delta_3 W_3$$

déterminée comme suit:

(1) On se donne une liste de désinences admissibles A_1 A_2.... Ces désinences sont choisies de façon à optimiser le temps de consultation moyen compte tenu des faux découpages (voir ci-dessus) δ_1 δ_2 sont les codes machine de ces désinences.

(2) Pour chacune des unités morphologiques d'une catégorie morphologique, on détermine la ou les bases *B* compte tenu des désinences de la liste précédente et *le paradigme linguistique* correspondant $\Pi = \Delta_1 W_1$, $\Delta_2 W_2$, ... Les désinences identiques d'un tel paradigme sont regroupées. Dans le paradigme réduit ainsi déterminé, on remplace les Δ et les *W* par leurs codes machines. Le paradigme obtenu est le paradigme-machine relatif à la base *B*.

La correspondance entre le paradigme linguistique et le paradigme-machine est établie à l'aide d'un simple numéro d'ordre (le numéro de paradigme) facile à noter et à perforer.

Les informations linguistiques communes aux diverses formes d'une même unité morphologique (en particulier la catégorie grammaticale) sont représentées par le code commun (C.C.) Une unité morphologique est ainsi représentée par B. Π (C.C.)

PROBLÈMES LINGUISTIQUES RELATIFS AUX PARADIGMES

Les paradigmes sont recensés systématiquement en examinant si toutes les combinaisons possibles des critères généraux relatifs aux déclinaisons et aux conjugaisons habituelles correspondent à des paradigmes susceptibles d'exister. Par exemple, pour le substantif russe: défectivité au singulier ou au pluriel, genre, déclinaison dure ou molle, animation etc. Les critères sont trop nombreux et parfois insuffisants pour être utilisés systématiquement en machine mais ils sont utiles pour dresser une première liste des paradigmes.

Les paradigmes irréguliers sont en général bien recensés par les grammaires et il suffit de les ajouter aux paradigmes déterminés par les critères généraux. Remarquons en particulier que ces paradigmes irréguliers sont traités en machine comme des paradigmes ordinaires et

qu'il est toujours possible de rajouter des paradigmes aux tables existantes sans modifier le programme lui-même.

Il est parfois nécessaire de considérer des paradigmes défectifs (par exemple, dans le cas d'une voyelle mobile pour un substantif russe). En général, la réunion de 2 paradigmes défectifs est elle-même un paradigme complet. Ceci permet de réduire le nombre de numéros de paradigmes en affectant à un paradigme complet, deux numéros de paradigme : ceux des paradigmes défectifs dont ce paradigme est la réunion.

La recherche d'un numéro de paradigme par le linguiste chargé de coder les mots d'un dictionnaire est facilitée par l'utilisation d'un *arbre de détermination* qui met en évidence les divers critères linguistiques généraux mentionnés ci-dessus.

RÉALISATION PRATIQUE

Le système précédent a été utilisé à la section de Paris du CETAP pour la réalisation d'un dictionnaire de bases de la langue russe, fonctionnant effectivement sur ordinateur 650 à disques 355.

Les mots russes sont classés en 5 catégories morphologiques :

(1) Substantifs et Pronoms personnels—variables grammaticales: cas et nombre—198 paradigmes—36 *W* ambigus.
(2) Adjectifs forme longue (adjectifs ordinaires, participes et adjectifs pronominaux) var. gram. : cas, genre, nombre et animation—33 paradigmes—22 *W* ambigus.
(3) Adjectifs forme courte—var. gram. : genre et nombre 24 paradigmes—7 *W* ambigus.
(4) Cardinaux—var. gram.: genre, nombre, cas—23 paradigmes—18 *W* ambigus.
(5) Verbes—var. gram.: temps, personne, nombre . . .—19 W ambigus—79 paradigmes pour la base du présent—18 paradigmes pour la base du passé.

Cette liste de paradigmes et de *W* peut être considérée comme étant pratiquement exhaustive. Ce système a permis d'indexer morphologiquement 12,000 bases russes environ. L'ensemble de ces bases couvre très largement un corpus d'environ 90,000 occurrences appartenant à des domaines très hétérogènes. Il couvre également la liste des mots russes les plus fréquents établis par H. Josselson.[5] Il a permis d'indexer automatiquement 15,000 formes environ d'un lexique de 23,000 formes établi par la Rand Corporation, bien que ce lexique soit celui d'un corpus homogène (physique).

L'indexage de ces 12,000 bases a été vérifié soigneusement par l'utilisation de 2 programmes auxiliaires: un programme de déclinaison automatique des mots restituant la liste des désinences d'une base dans l'ordre habituel aux linguistes et un programme de décodage des *W* permettant de restituer en clair les informations relatives au *W* que l'on peut associer à une forme donnée.

BIBLIOGRAPHIE

[1] A. G. OETTINGER: *Automatic Language Translation.* Harvard University Press (1960).
[2] H. S. KELLY and T. W. ZIEHE: *Glossary Look up Made Easy.* Rand P-1909 (1960).
[3] S. M. LAMB and W. H. JACOBSEN: A high-speed large capacity dictionary system. J. *Mech. Transl. M.I.T.* (Nov. 1961).
[4] D. W. DAVIES and A. M. DAY: A technique for consistent splitting of Russian words. 1ère Conférence Internationale sur la Traduction Automatique, Londres (September 1961).
[5] H. H. JOSSELSON: *The Russian Word Count.* Detroit Wayne University Press (1953).
[6] P. MEILE: On problems of address in an automatic dictionary of french—1ère Conférence Internationale sur la Traduction Automatique, Londres (September, 1961).

[7] G. VEILLON: Consultation d'un dictionnaire et analyse morphologique en Traduction Automatique. Thèse présentée à l'Université de Grenoble (1962).
[8] M. GROSS: On the equivalence of models of language used in the fields of mechanical translation and information retrieval. Venise (Juillet 1962).
[9] A. SESTIER: Un système de dictionnaire automatique. 2ème Congrès de l'AFCALTI. Note CETAP No. 11 Paris (1961).
[10] L. DUPUIS : Un système d'analyse morphologique de la langue russe. 2ème Congrès de l'AFCALTI, Note CETAP No. 12, Paris (1961).
[11] A. SESTIER: Sur la nécessité et la possibilité des normalisations dans les échanges de dictionnaires en Traduction Automatique. Colloque de l'Association pour la Traduction Automatique et la Linguistique Appliquée, Louveciennes (Mars 1962).
[12] Note CETAP No. 14. L'outillage d'Etude du Russe au CETAP (1962).
[13] B. DUPUIS: Analyse morphologique de la Langue Allemande. Note CETAP No. 9 (1962).
[14] G. O. TAILLEUR: Les Paradigmes du Français. Note CETAP No. 16 (1962).