Contribution à la synthèse automatique du français

Michel Berthaud

N° d'ordre

THESES

présentées à

LA FACULTE DES SCIENCES DE L'UNIVERSITE DE GRENOBLE

pour obtenir

LE GRADE DE DOCTEUR - INGENIEUR

---

par

Michel BERTHAUD
Ingénieur A.M., I.M.A.G.

Première thèse :

CONTRIBUTION A LA SYNTHESE AUTOMATIQUE DU FRANCAIS

Deuxième thèse :

PROPOSITIONS DONNEES PAR LA FACULTE

Thèses soutenues le octobre 1964 devant la Commission d'examen :

Monsieur J. KUNTZMANN, Président
Messieurs B. VAUQUOIS , Examinateurs
N. GASTINEL

PROFESSEURS SANS CHAIRE

M.	LACASE A.	THERMODYNAMIQUE
Mme	KOFLER L.	BOTANIQUE
MM.	DREYFUS B.	THERMODYNAMIQUE
	VAILLANT F.	ZOOLOGIE ET HYDROBIOLOGIE
	GIRAUD P.	GEOLOGIE
	GIDON P.	GEOLOGIE ET MINERALOGIE
	ARNAUD P.	CHIMIE
	PERRET R.	SERVOMECANISMES
Mme	LUMER L.	MATHEMATIQUES
Mme	BARBIER M.J.	ELECTROCHIMIE
Mme	SOUTIF J.	PHYSIQUE
MM.	BRISSONNEAU P.	PHYSIQUE
	COHEN J.	ELECTROCHIMIE
	DEPASSEL R.	MECANIQUE
	GASTINEL N.	MATHEMATIQUES APPLIQUEES

PROFESSEURS ASSOCIES

MM.	LUMER G.	MATHEMATIQUES
	HIGUCHI	BIOSYNTHESE DE LA CELLULOSE
	WAGNER	BOTANIQUE

MAITRES DE CONFERENCES

MM.	ROBERT A.	CHIMIE PAPETIERE
	ANGLES D'AURIAC	MECANIQUE DES FLUIDES
	BIAREZ J. P.	MECANIQUE PHYSIQUE
	COUMES A.	ELECTRONIQUE
	DODU J.	MECANIQUE DES FLUIDES
	DUCROS P.	MINERALOGIE ET CRISTALLOGRAPHIE
	CLENAT P.	CHIMIE
	HACQUES G.	CALCUL NUMERIQUE
	LANCIA R.	PHYSIQUE AUTOMATIQUE
	PEBAY-PEROULA	PHYSIQUE
	KAHANE	PHYSIQUE GENERALE
	DOLIQUE	ELECTRONIQUE
Mme	KAHANE J.	PHYSIQUE
MM.	DEGRANGE C.	ZOOLOGIE
	GAGNAIRE D.	CHIMIE PAPETIERE
	RASSAT A.	CHIMIE SYSTEMATIQUE
	KLEIN J.	MATHEMATIQUES
	BETHOUX P.	MATHEMATIQUES APPLIQUEES
	POULOUJADOFF M.	ELECTROTECHNIQUE
	DEPOMMIER P.	PHYSIQUE NUCLEAIRE
	DEPORTES C.	CHIMIE
	BARRA J.	MATHEMATIQUES APPLIQUEES
Mme	BOUCHE L.	MATHEMATIQUES
MM.	PERRIAUX J.	GEOLOGIE
	SARROT-REYNAULD	GEOLOGIE
	CAUQUIS G.	CHIMIE GENERALE
	LABBE A.	BOTANIQUE
	BONNET G.	PHYSIQUE GENERALE
	BARNOUD F.	BIOSYNTHESE DE LA CELLULOSE
Mme	BONNIER M. J.	CHIMIE

MAITRES DE CONFERENCES ASSOCIES

MM.	ISHIKAWA Y.	MAGNETISME
	QUATTROPANI	THERMODYNAMIQUE

FACULTE DES SCIENCES

L I S T E D E S P R O F E S S E U R S

DOYENS HONORAIRES

M. FORTRAT P.

M. MORET L.

DOYEN

M. WEIL L.

PROFESSEURS TITULAIRES

MM.	NEEL L.	MAGNETISME ET PHYSIQUE DU SOLIDE
	DORIER A.	ZOOLOGIE
	HEILMANN R.	CHIMIE ORGANIQUE
	KRAVTCHENKO J.	MECANIQUE RATIONNELLE
	CHABAUTY C.	CALCUL DIFFERENTIEL ET INTEGRAL
	PARDE M.	POTAMOLOGIE
	BENOIT J.	RADIOELECTRICITE
	CHENE M.	CHIMIE PAPETIERE
	BESSON J.	ELECTROCHIMIE
	WEIL L.	THERMODYNAMIQUE
	FELICI N.	ELECTROSTATIQUE
	KUNTZMANN J.	MATHEMATIQUES APPLIQUEES
	BARBIER R.	GEOLOGIE APPLIQUEE
	SANTON L.	MECANIQUE DES FLUIDES
	OZENDA P.	BOTANIQUE
	FALLOT M.	PHYSIQUE INDUSTRIELLE
	GALVANI O.	MATHEMATIQUES
	MOUSSA A.	CHIMIE NUCLEAIRE
	TRAYNARD P.	CHIMIE
	SOUTIF M.	PHYSIQUE
	CRAYA A.	HYDRODYNAMIQUE
	REULOS R.	THEORIE DES CHAMPS
	AYANT Y.	PHYSIQUE APPROFONDIE
	GALLISSOT F.	MATHEMATIQUES APPLIQUEES
Melle	LUTZ E.	MATHEMATIQUES
MM.	BLAMBERT M.	MATHEMATIQUES
	BOUCHEZ R.	PHYSIQUE NUCLEAIRE
	ILLIBOUTRY L.	GEOPHYSIQUE
	MICHEL R.	GEOLOGIE ET MINERALOGIE
	BONNIER E.	ELECTROCHIMIE
	DESSAUX G.	PHYSIQUE ANIMALE
	PILLET E.	ELECTROCHIMIE
	DEBELMAS J.	GEOLOGIE
	GERBER R.	MATHEMATIQUES
	PAUTHENET R.	ELECTROTECHNIQUE
	VAUQUOIS B.	MATHEMATIQUES APPLIQUEES
	SILBER R.	MECANIQUE DES FLUIDES
	MOUSSIEGT J.	ELECTRONIQUE
	BARBIER J. C.	PHYSIQUE
	KPSZUL J. L.	MATHEMATIQUES
	BUYLE-BODIN M.	ELECTRONIQUE

Je tiens à exprimer ma profonde reconnaissance

A Monsieur le Professeur KUNTZMANN,
Directeur de l'Institut de Mathématiques Appliquées de Grenoble, qui a bien voulu me faire l'honneur de présider le Jury de thèse,

A Monsieur le Professeur VAUQUOIS,
Directeur du Centre d'Etudes pour la Traduction Automatique, qui a dirigé ce travail et qui par son aide et ses conseils judicieux m'a permis de le mener à bien,

A Monsieur le Professeur GASTINEL, qui a bien voulu faire partie du Jury.

Je remercie Monsieur le Professeur de POSSEL,
Directeur de l'Institut Blaise Pascal de Paris, d'avoir parrainé mes recherches dans le cadre du Centre National de la Recherche Scientifique.

Que les membres du Centre d'Etudes pour la Traduction Automatique, en particulier Messieurs VEYRUNES et VEILLON, trouvent ici l'expression de ma gratitude pour l'aide efficace qu'ils m'ont apportée.

Je remercie également les membres du Laboratoire de Calcul, en particulier Monsieur MOUNET, qui ont permis la réalisation matérielle de cet ouvrage.

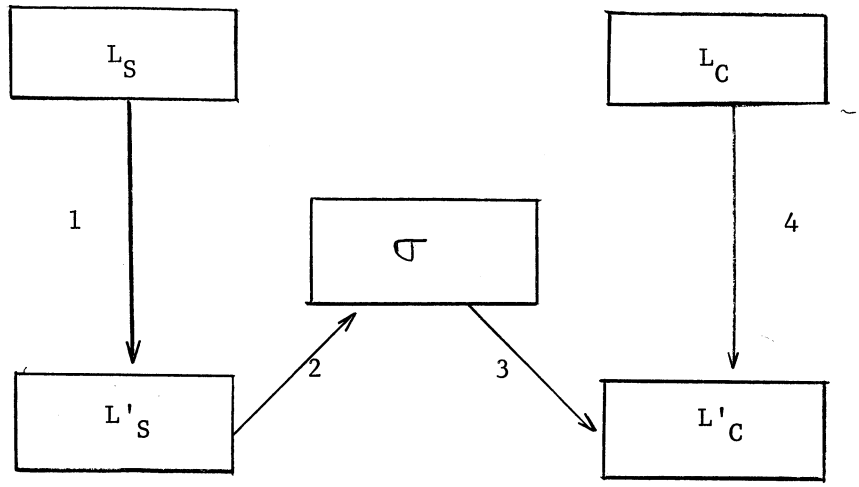
CONTRIBUTION A LA SYNTHESE AUTOMATIQUE

DU FRANCAIS

INTRODUCTION

Les premiers essais de Traduction Automatique ont utilisé le "mot à mot" ; par ce procédé chaque mot du texte source est remplacé par le mot correspondant en langue cible, à l'aide d'un dictionnaire bilingue. Plusieurs inconvénients découlent de l'emploi de cette méthode : les cas d'homonymies et les problèmes syntaxiques ou sémantiques inhérents à chaque langue ne sont pas traités. Par suite, une étude plus complète des langues a semblé nécessaire : formation des mots, construction des phrases et sens attachés à chaque forme.

En vue de l'utilisation de ces données linguistiques sur ordinateur et dans un but de précision, on est arrivé à la représentation des langages naturels à l'aide de systèmes formels . Ceux-ci ont été étudiés et font appel à des théories connues (machine de Turing, grammaires, ...). Entre deux langues quelconques la communication est assurée par un ensemble commun de concepts qui figurent les idées ; à un sous-ensemble de concepts correspondent des formulations différentes dans les langues distinctes, la pensée exprimée étant la même. Cet ensemble de concepts sera appelé langage intermédiaire (ou langage-pivot) [1] . Le schéma suivant distingue les phases successives d'un procédé de Traduction Automatique (le cheminement est indiqué par des flèches).



L'_S est le modèle formel du langage naturel L_S (source). La formalisation (1) est établie selon certaines règles et L'_S est décrit à l'aide d'un métalangage de base. Le processus est identique pour la langue cible L_C (4), les contraintes de formalisation étant adaptées à la synthèse et non à l'analyse. Ainsi L'_C est un sous-ensemble de L_C (en particulier les synonymes de L_C conduisent à un élément unique de L'_C). En résumé, L'_S doit être capable de reconnaître toutes les formes et toutes les structures syntaxiques valables sur L_S, alors que le vocabulaire et la grammaire de L'_C peuvent être limités volontairement, sans pour autant diminuer la puissance d'expression de L_C .

La partie proprement automatique de la traduction correspond au passage entre les deux modèles formels L'_S et L'_C. Par l'intermédiaire de 2 le texte à analyser est représenté par un sous-ensemble de concepts sur le langage-pivot σ, dont la figuration sur L'_C est assurée par 3, selon les règles propres à la

langue cible.

La définition et la composition de σ ne pourront se déduire que d'une étude sémantique précise des différentes langues naturelles. A priori, le contenu de ne peut être qu'intuitif. Actuellement, nous ne ferons appel aux notions sémantiques que pour poser certains problèmes, étant bien entendu que la possibilité d'un passage par σ est toujours envisagée.

Dans ce qui suit, nous essairerons de fixer un cadre au problème de la synthèse d'une langue. Quelques réalisations pratiques donneront un aperçu des résultats obtenus, et nous préciserons les phases qui nécessitent une étude plus approfondie.

CHAPITRE I

RAPPELS SUR L'ANALYSE D'UNE LANGUE

I - ANALYSE MORPHOLOGIQUE DE LA LANGUE SOURCE

On appellera "forme" un terme construit à l'aide du vocabulaire V_S et appartenant à L_S . Le texte à traduire est composé d'une chaîne d'occurrences assemblée selon les règles de construction valables pour L_S . L'occurrence est l'apparition de la forme (suite de lettres comprise entre deux blancs ou deux séparateurs).

Selon le principe adopté, chaque forme est décomposable en trois élements principaux :

<forme> ::= <préfixe> <base> <affixe>

L'affixe peut être aussi obtenu par concaténation de plusieurs parties (suffixe, désinence, suffixe de réflexivité). L'ensemble des formes générées (ou reconnues) à partir d'une base, appartenant à une classe morphologique donnée, est appelé unité lexicale ou U.L. [1, pages 37 à 39] . Par exemple, pour un verbe toutes les formes obtenues par conjugaison aux temps simples constituent la même unité lexicale.

Pour chaque classe morphologique, il existe un certain nombre de variables grammaticales (par exemple, substantif : genre, nombre, ...). La base d'une part, l'affixe d'autre part déterminent certaines valeurs de ces variables. Une dérivation consiste en un pré-

fixe ou un affixe (ou les deux) ; elle fait passer d'une classe morphologique à une catégorie lexicale (syntaxique), et peut avoir les effets suivants :

- Modification des noms des variables grammaticales
- Modification de la valeur sémantique attachée à l'unité lexicale (par exemple, négation).

Les dérivations sont spécifiques à chaque classe morphologique.

Par exemple, les formes suivantes :

ORDRE
ORDRES
DESORDRE
DESORDRES

appartiennent à la même U.L. , comportant la base unique ORDRE .

La définition ci-dessus est purement formelle et la pratique a introduit un critère de rentabilité. Un affixe (ou un préfixe) sera pris en considération s'il permet le découpage d'un grand nombre de formes.

Dans le cas contraire, ces flexions ne seront pas retenues ; par suite l'unité lexicale comportera plusieurs bases, chacune étant défective (cas d'un substantif ayant un pluriel irrégulier). De même l'unité lexicale est défective lorsqu'elle n'admet pas toutes les flexions attachées à la classe morphologique à laquelle elle appartient.

Les informations morphologiques ainsi définies sont codées et rassemblées dans le syntagme élémentaire (composant de base de L'_S). Certaines valeurs de ces informations sont statiques - lues dans un dictionnaire - , d'autres sont dynamiques - calculées selon la désinence et la dérivation.

$\langle$s.e.$\rangle$ ::= $\langle V \rangle$ $\langle$n° identificateur$\rangle$ $\langle K_{u,v} \rangle$ $\langle vg_p \rangle$ $\langle vg_c \rangle$ $\langle$c.s.$\rangle$

V est le numéro d'ordre de l'occurrence dans le texte (n° séquentiel).

<n° identificateur> ::= <CC> <n° U.L.> <n° de base>

CC est le code classe (voir en particulier [11]), n° U.L. définit le numéro d'unité lexicale et le n° de base précise de quelle base il s'agit dans l'U.L. . Cette valeur est donnée par le dictionnaire.

$K_{u,v}$	Catégorie syntaxique
Δ	Numéro de dérivation
vg_p	1 valeur de variable grammaticale permanente (fournie par l'U.L.)
vg_c	1 valeur de variable grammaticale contingente (calculée à l'aide de la désinence)
c.s.	Code syntaxique (recopié à partir du dictionnaire)

Ces définitions ont été adoptées au CETA [1, 9, 10 et 11] .

II - ANALYSE SYNTAXIQUE ET INTERPRETATION

L'analyse morphologique n'est pas capable d'éliminer les homographes : à une occurrence correspond, en général, plusieurs syntagmes élémentaires. Il est nécessaire de procéder à une analyse syntaxique du texte (ou plus simplement de la phrase) de manière à reconnaître les dépendances syntaxiques qui existent entre les s.e.. Un certain nombre d'homographes seront supprimés du fait des accords grammaticaux impliqués par des règles de grammaire. L'introduction

d'un code sémantique permettra de choisir entre plusieurs structures lorsque la phase est syntaxiquement ambigüe : nous rejoignons ici le niveau du langage - pivot σ [12, 13 et 24] .

A - Grammaire de reconnaissance

(Pour les définitions, voir en particulier [16]).

Un langage L et une grammaire G sont définis sur $V = V_A \cup V_T$

V_T vocabulaire terminal : ensemble fini de mots a_i .

V_A vocabulaire auxiliaire : ensemble fini des éléments intervenant dans l'écriture des règles de grammaire, V_A contient un symbole distingué P qui est l'axiome de la grammaire.

Utilisée en génération, une règle s'écrit sous la forme générale

$$(R) \qquad \varphi \longrightarrow \Psi \qquad (\varphi \text{ est réécrit } \Psi)$$

Pour une grammaire de type context-free , φ est un symbole unique $A \in V_A$ et Ψ est une séquence, non vide, sur V . Cette classe de langages peut être décrite à l'aide d'un automate à mémoire "pushdown".

En reconnaissance, les règles utilisées ont été mises sous la forme suivante :

$$\begin{array}{l} a \Longrightarrow A \\ BC \Longrightarrow A \end{array} \qquad \text{avec} \begin{cases} a \in V_T \\ A, B, C \in V_A \end{cases}$$

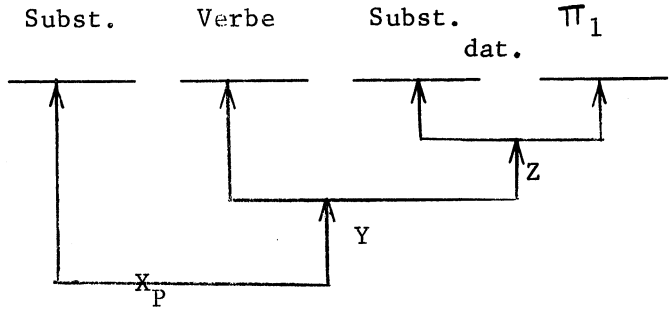
dérivée de la forme normale de CHOMSKY.

L'arbre de représentation de la structure syntaxique sera de mode binaire. Les mots du vocabulaire terminal sont les syntagmes élémentaires ; les règles de la forme $a \Longrightarrow A$ sont appliquées lors d'une phase pré-syntaxique, après la morphologie. V_A est aussi appelé vocabulaire non terminal, ses éléments sont les syntagmes (S_i) qui sont définis par les linguistes selon la langue étudiée et les règles de construction adoptées. Le code syntaxique d'un syntagme permet de choisir la règle adéquate parmi le sous-ensemble de règles où ce S_i apparait en partie gauche.

Certaines règles, au moment de leur utilisation, saturent un (ou deux) élément de partie gauche d'une (ou plusieurs) autre règle ; de ce fait, le choix d'une règle dépend de celles utilisées à un niveau inférieur dans le même sous-arbre : cette méthode réduit le nombre des éléments non terminaux (donc des règles). La grammaire utilisée est intrinséquement continue ; cependant elle est capable de traiter certains cas de constituants discontinus en leur associant une structure syntaxique qui ne rend pas compte des liaisons effectives.

EXEMPLE

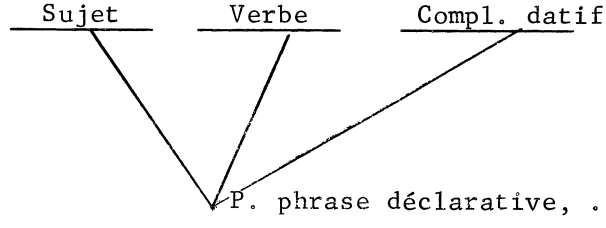
En allemand, la particule séparable modifie le sens d'un verbe ainsi que les valeurs des variables syntaxiques (nouvelles possibilités d'accord) : un verbe n'admettant pas de complément au datif sera modifié par la présence d'une particule qui, elle, l'autorisera.



La structure fournie ne décrit pas les effets exacts de la particule ; l'interprétation viendra modifier les liaisons.

B - Rôle de l'interprétation

La structure syntaxique déterminée sert de support à une interprétation de la phrase. Le langage interprétatif est formé d'un vocabulaire, mais on n'est pas libre du vocabulaire auxiliaire (non terminal) ; en effet, les éléments sont précisés à partir de la structure syntaxique. Ainsi, certaines valeurs de variables grammaticales et syntaxiques sont "remontées" dans l'arbre au niveau le plus élevé : on pourra distinguer phrase déclarative, sujet animé, complément d'agent, etc ... Dans d'autres cas, la structure est "redistribuée" sous une forme qui peut être non binaire. Ainsi l'exemple précédent devient :



L'interprétation doit donc permettre de faire ressortir la composition de la phrase avec les fonctions de chacun des éléments. Cette étape est à mi-chemin de la sémantique qui permettrait de trouver la signification qui est l'invariant lors de la traduction ; actuellement, nous ne considérerons pas de renseignements sémantiques pour lever les ambiguïtés qui pourraient exister lors de la construction de la structure interprétée : on parlera de variables "interprétatives" .

On peut remarquer cependant que les éléments du vocabulaire du langage d'interprétation peuvent être définis en fonction de

la langue cible, puisqu'elle est le point commun des langues analysées, ce qui a pour conséquence : il n'y a pas lieu de résoudre une ambiguïté de L_S au niveau interprétatif, si cette ambiguïté se conserve dans L_C .

Au point de vue stratégie, la reconnaissance des structures syntaxiques et leur interprétation peuvent constituer deux phases distinctes, ou n'en former qu'une seule. En général, une phrase de L_S donnera naissance à plusieurs structures interprétées P_1 , P_2 , ... P_n , chacune d'entre elles fournissant une seule structure en langue cible : on peut dire que c'est le but assigné à l'interprétation.

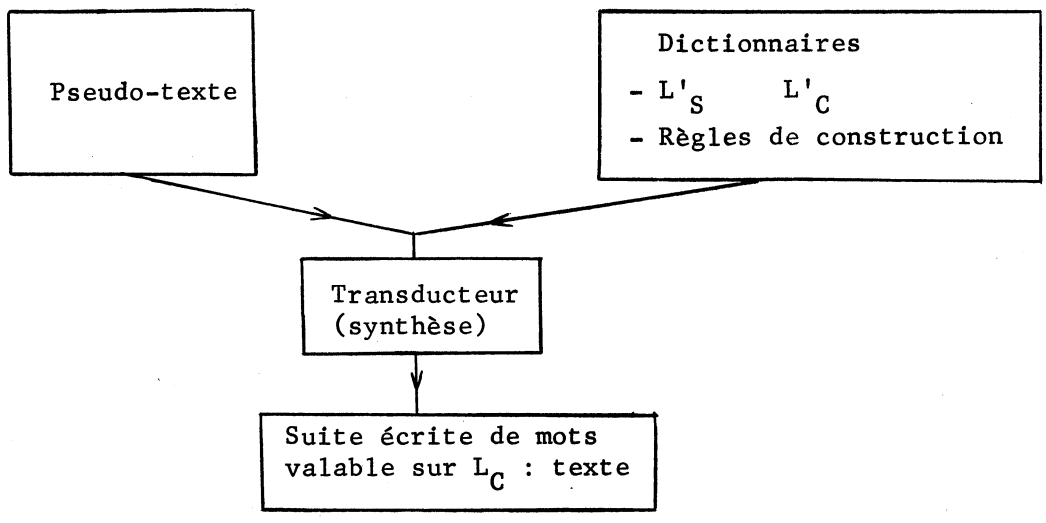
CHAPITRE II

ORGANISATION GENERALE DE LA SYNTHESE D'UNE LANGUE

I - BUT ET MOYENS DE LA SYNTHESE

L'élément de traitement est la structure interprétée qui est fournie par l'analyse de la langue origine. Pour effectuer une traduction correcte, il est nécessaire de modifier la structure en fonction de la langue cible choisie, ici le Français. Nous appellerons "pseudo-texte" la chaîne codifiée qui sert de point de départ.

La chaîne de sortie est un "texte" en français. Il faut donc construire un "transducteur" effectuant le passage d'informations codées en une suite de mots compréhensible directement à la lecture. Des études linguistiques sont nécessaires pour mener à bien ce travail : dictionnaire bilingue, renseignements syntaxiques et morphologiques. En particulier, il faut préciser le vocabulaire acceptable en langue cible et déterminer les équivalents selon les termes de la langue source.



II - GRAMMAIRE DE TRANSFERT

A - Nécessité

L'intérêt d'une traduction mot à mot est très limité. Prenons l'exemple proposé par Klima et repris par Gross [15] :

Soit la phrase anglaise (1) qui traduite en français mot à mot donne (1')

He dwelled on its advantages . (1)

Il a insisté sur ses avantages. (1')

Considérons la forme passive (2) de (1)

Its advantages were dwelled on by him. (2)

En français (1') est aussi la traduction de (2) car il n'y a pas de forme passive pour cette phrase. Pour aboutir à ce résultat, deux méthodes sont disponibles :

- Transformation de (2) en (2') qui n'est pas une phrase française.
  Ses avantages ont été insisté sur par lui (2')
  Puis passage de (2') en (1').
- Transformation de (2) en (1) avant traduction.

Cette seconde méthode semble plus naturelle. Chaque phrase sera donc donnée par sa structure, et une grammaire de transfert change les arbres d'entrée en arbres de sortie. D'un point devue formel, une grammaire de transfert est précisément une grammaire transformationnelle. Cette notion, introduite par Chomsky, fait actuellement l'objet des travaux de Matthews [15].

B - Traducteur fini

Soit un automate à mémoire pushdown possédant un nombre fini d'états. La phrase proposée p est sur le ruban d'entrée ; à l'état stable (ou final) la mémoire pushdown contient une phrase p' dont les mots sont des éléments du vocabulaire de sortie.

Ce système traduit une phrase d'entrée p en une phrase de sortie p' : il est appelé traducteur fini. Soit L un langage dont les phrases p sont proposées à l'entrée d'un traducteur T ; si, pour tout $p \in L$, T traduit p en p' on désigne par T (L) l'ensemble des phrases p' .

Si les règles de T peuvent être mises sous la forme :

$$(a_i , S_j) \longrightarrow (S_k , A_e)$$

S_j et S_k sont des états internes du traducteur,

$a_i \in$ vocabulaire d'entrée, $A_e \in$ vocabulaire de sortie.

Alors T est un traducteur borné. En particulier, on n'aura pas de boucles fermées d'instructions (X , S_j) (S_k , A_e) où l'application de la règle est indépendante du mot lu et n'entraine pas de déplacement du ruban d'entrée.

Pour la traduction automatique, le schéma suivant proposé par Ingve est couramment adopté : étant donnés deux langages L et L' décrits respectivement par les grammaires G et G' indépendantes, on associe une grammaire de transfert T qui assure la communication L vers L' , ou L' vers L [4] . On voit ainsi que cette grammaire peut être représentée par un traducteur. On démontre (Ginsburg et Rose) que, L et L' étant des langages context-free, il n'y a pas d'algorithme qui détermine s'il existe un traducteur fini T , tel que $T(L) = L'$.

Pour que T soit acceptable, il faut que $T(L) \subset L'$; de plus, le traducteur ne doit pas réduire la généralité du langage L', ce qui signifie que les phrases de $L' \notin T(L)$ sont des particularités (de style, par exemple) de phrases appartenant à $T(L)$. Remarquons que la grammaire de transfert dépend dans ce cas des langues L et L' .

C - Transfert à partir d'une structure interprétée

Dans le paragraphe précédent, la phase interprétation n'est pas considérée : il y a transfert aux niveaux syntaxiques. Nous supposerons, maintenant, que la structure dont nous disposons a été interprétée en fonction de la langue cible, c'est à dire que les éléments terminaux et non-terminaux sont caractérisés par des noms et des valeurs de variables ayant une signification relativement à la langue cible.

1) - Mode de représentation :

Nous avons vu que la structure syntaxique peut être représentée par un arbre. Plus généralement, nous dirons que la structure interprétée est figurée par un graphe orienté ne contenant pas de circuits ; il peut donc, y avoir plusieurs branches "entrant" dans un noeud [8] . Une racine est un noeud dans lequel il n'y a pas de branche "entrant" ; s'il n'existe pas de branche "sortant", c'est une feuille (ou noeud terminal). Un noeud quelconque définit un sous-graphe, dont il est la racine, formé de ce noeud et de tous les noeuds atteints à partir de celui-ci avec les mêmes associations que le graphe principal.

Dans ce cas, chaque noeud représente un élément du vocabulaire appartenant au langage d'interprétation. La quantité d'information attachée à chaque noeud est décomposable en deux parties : le nom et la fonction.La fonction indique les valeurs des différentes variables (sémantiques, syntaxiques ou morphologiques) qui sont attachées au nom.

2) - Opérations de base en transfert de structures :

Les opérations élémentaires suivantes suffisent pour décrire une transformation [21 et 6] .

- Insertion d'un ou plusieurs noeuds avec leurs associations
- Suppression d'un sous-graphe
- Modification d'un ou plusieurs noeuds
- Permutation ou réarrangement des noeuds.

En traduction Russe - Français, l'insertion est un fait fréquent : article devant un substantif, prépositions pour in-

troduire les compléments, etc ... En Allemand, une particule séparable peut être associée à un verbe dont elle modifie les caractéristiques, une transformation supprime cette particule qui n'a pas d'existence en français. La modification intervient, par exemple, si le temps du verbe d'une proposition subordonnée dépend du temps du verbe de la proposition principale, les règles d'accord n'étant pas les mêmes dans les deux langues. Nous aurons enfin permutation ou réarrangement dans le cas où l'ordre adjectif - nom est inversé ; cela se produit aussi lorsque l'ordre des compléments dépend de leur type.

En général, une transformation nécessitera plusieurs de ces opérations élémentaires. Plusieurs transformations peuvent être appliquées successivement à une même partie de la structure ; l'ordre dans lequel celles-ci interviennent ne peut être arbitraire. On peut essayer d'ordonner ces transformations, qui correspondent chacune à une règle de transfert, selon le sous-graphe sur lequel elles opèrent (sujet, groupe verbal, phrase, etc ...) et selon le type d'opérations élémentaires qui les composent. Ceci est impossible car chaque transformation n'utilise pas une seule opération. Après une étude des règles existantes pour la traduction du Russe vers l'Anglais, Oettinger [21] a adopté la classification suivante des règles, selon les opérations effectuées : suppression, insertion, modification et réarrangement. La grammaire de transfert sera donc séquentielle dans le sens où il y a une hiérarchie dans l'application des règles.

En fonction de ces principes généraux, une règle peut s'écrire de la manière suivante :

$$\varphi\, f\, \Psi \longrightarrow \varphi\, f'\, \Psi$$

où φ, f , Ψ et f' sont des séquences, non toutes vides, appartenant au langage d'interprétation.

Cas particuliers :

a) $f' = \emptyset$ (vide) suppression. Dans ce cas, et peuvent être vides

b) $f = \emptyset$ avec $f' \neq \emptyset$ insertion selon les dépendances

c) f se réduit à un seul élément A :

- f' est une séquence de longueur supérieure à 1 contenant A : insertion.
- f' = A' : modification.

Le cas général décrit les règles utilisant toutes les opérations de base.

D - Quelques difficultés du transfert de structure

Dans les langues à déclinaisons, la fonction syntaxique se représente par une désinence, ce qui sera traduit en français par une préposition. Prenons les deux exemples suivants :

Je vais en Bourgogne (1)
Je vais à Paris (2)

" Bourgogne" et "Paris" sont complément de lieu, la destination étant marquée par "à" ou "en" . Il est possible de les distinguer en remarquant que "Bourgogne" est une région et "Paris" une ville, ce qui nécessite un code adjoint à ces mots qui permet de faire un choix sur la préposition.

Pour les phrases (3) et (4) , une règle permet de résoudre ce problème :

Je vais en Allemagne (3)

Je vais au Maroc (4)

Car "Allemagne" est du genre féminin, "Maroc" étant du genre masculin.

Par contre, pour (5) et (6) on ne peut invoquer qu'une règle phonétique qu'il est impossible de systématiser

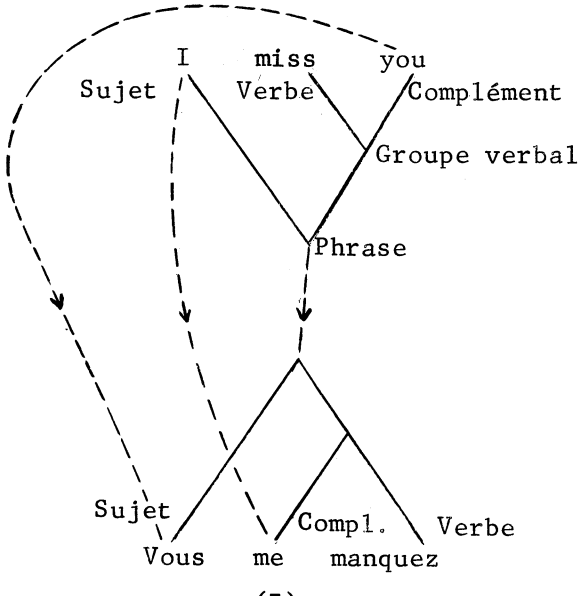
Nous sommes au printemps (5)

Nous sommes en hiver (6)

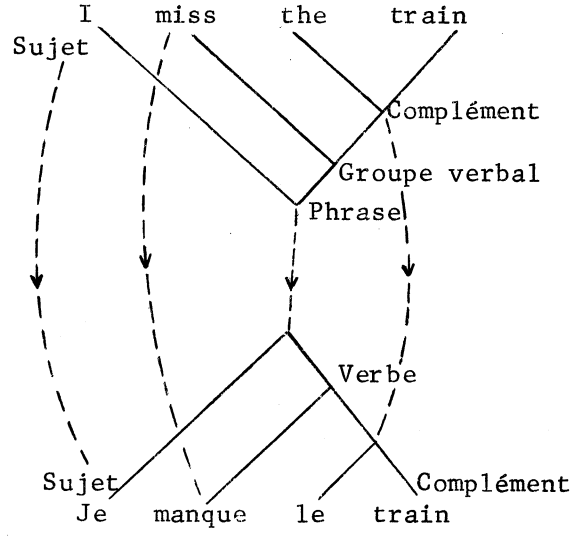
D'après les exemples précédents, on peut remarquer que le choix de la préposition est fait à différents niveaux :

- Pour (1) et (2) ce sont des considérations sémantiques (qui peuvent être ramenées au niveau syntaxique) qui permettent de résoudre le problème.
- Pour (3) et (4) on tiendra compte des accords grammaticaux.
- Enfin, pour (5) et (6) la décision intervient au niveau phonétique.

Pour terminer ce bref survol des difficultés posées par le transfert de structure, nous citerons l'exemple donné par Gross [15].



(7)



(8)

Les deux phrases ont des structures identiques en anglais ; en français, les fonctions des mots ont changé (le complément est devenu sujet, etc ...) pour (7). Les considérations invoquées pour les exemples précédents ne jouent aucun rôle, à priori. Pour traduire correctement la phrase (7) on peut avancer l'une des deux hypothèses suivantes :

- (7) est une locution figée en anglais (de même que "of course").
- si le complément du verbe "miss" est animé (personne) alors la structure équivalente en français subit une transformation. Remarquons que le verbe " manquer" est intransitif dans (7) et transitif dans (8).

Rappelons que la traduction correcte de l'exemple donné dans ce chapitre au paragraphe II - A fait intervenir des considérations syntaxiques relatives à l'intransitivité du verbe équivalent en français.

III - STRATEGIE DE LA SYNTHESE EN TRADUCTION AUTOMATIQUE

L'étude précédente permet de dégager un certain nombre d'étapes dans la partie synthèse :

- Obtention d'une structure valable pour la langue cible.
- Exploitation des liaisons, c'est à dire accords grammaticaux imposés par la syntaxe.
- Synthèse du texte, c'est à dire concaténation des mots pour obtenir la phrase sous une forme compréhensible à la lecture humaine.

Nous allons reprendre ces points en détail de manière à établir une suite cohérente et logique, en notant les incidences linguistiques et machine qui en découlent.

A - Transfert de structure

Une transformation agit, en général, sur plusieurs unités syntaxiques ; de plus, le transfert nécessite divers renseignements d'ordre sémantique ou syntaxique (voir paragraphe II - D). Avant toute modification de la structure, il faudra donc obtenir les codes relatifs aux "feuilles" du graphe représentatif. Le code sémantique sera conçu de manière à permettre le choix entre les équivalents possibles et le code syntaxique détermine les conditions d'application des règles de grammaire. Ce sera en fait une traduction "mot à mot " avec la différence essentielle qu'un certain nombre d'homographes intermes ont été éliminés et que chaque noeud est "interprété" c'est à dire qu'il a une "fonction".

A partir de ces éléments et de la structure existante il est alors possible de déterminer dans l'ensemble des structures équivalentes celle qui traduit la phrase interprétée.

Etant donné l'encombrement en machine d'un tel dictionnaire d'équivalents, nous verrons qu'il est impossible de fournir en même temps le code morphologique. Certains éléments "insérés" ne pourront pas être précisés (voir phrases (3) et (4) du paragraphe II-D) et un sous-ensemble de formes leur sera associé ; le "nom" de ces noeuds portera seulement l'indication d'une catégorie lexicale. Nous avons remarqué que ce cas ne se présentait que pour les classes fermées (articles, prépositions, etc ...).

B - Accords grammaticaux

Nous devons disposer de certains renseignements morphologiques associés à chaque "feuille" ; selon les liaisons et les valeurs des variables grammaticales nous pouvons compléter la détermination de certains éléments qui peuvent avoir plusieurs formes (par exemple adjectifs, verbes, etc ...).

C - Synthèse morphologique

Nous avons un dictionnaire contenant les suites de lettres qui donneront après concaténation les formes appartenant à la langue cible, donc les phrases. Certains problèmes annexes d'ordre phonétique (élision, contraction) et spécifiques au français seront traités, ainsi que l'édition du texte traduit.

CHAPITRE III

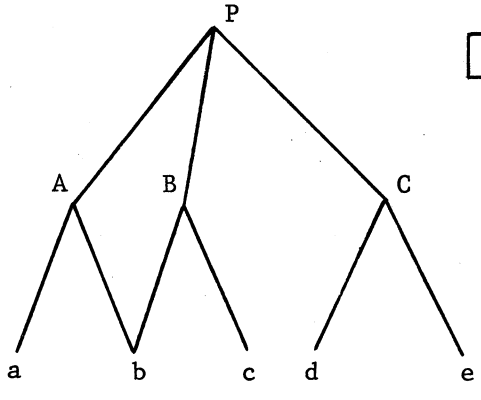
TRANSFERT DE STRUCTURE

ACCORDS GRAMMATICAUX

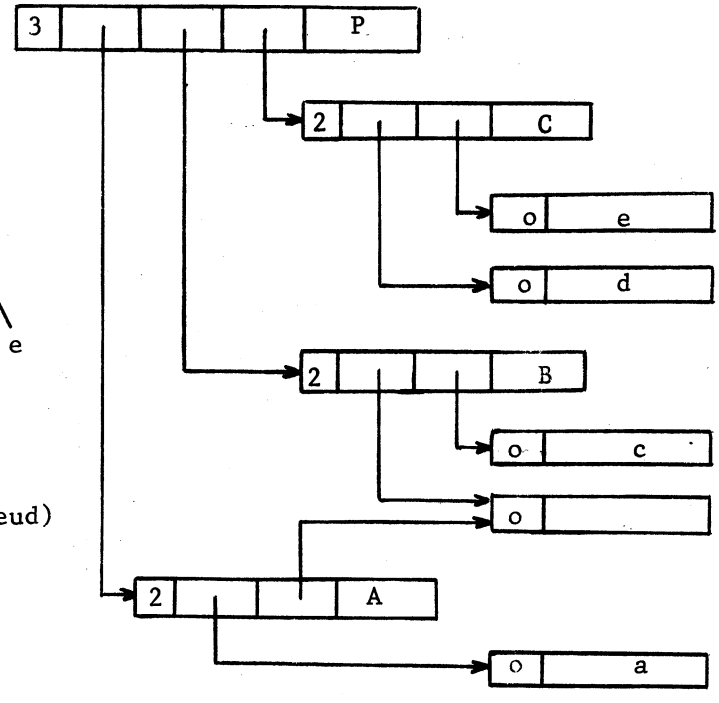
Cette phase assure la jonction entre la langue source et la langue cible, l'interprétation étant l'invariant dans la traduction ; par suite, nous sommes encore dépendant de la langue origine.

I - FORME MACHINE DE LA STRUCTURE INTERPRETEE

Un arbre peut se représenter à l'aide d'une liste, on peut supposer qu'il en sera de même pour les graphes particuliers que nous considérons.



Le nombre de successeurs (branches partant d'un noeud) est indiqué en tête .



Cette disposition n'est pas rigide : on pourrait, en particulier, supposer que l'adresse du noeud successeur le plus à gauche est implicite et qu'il est placé immédiatement après le noeud origine [14] . On peut remarquer que le fait d'avoir des adresses explicites conserve l'ordre normal des feuilles. Cette représentation permet une exploration très rapide du graphe, à partir de la racine P ; de plus, il est facile de supprimer ou d'insérer un sous-graphe. Par contre, étant donnée une feuille il est impossible de déterminer ses prédécesseurs ou de décider si elle appartient à un sous-graphe donné sans balayer entièrement celui-ci.

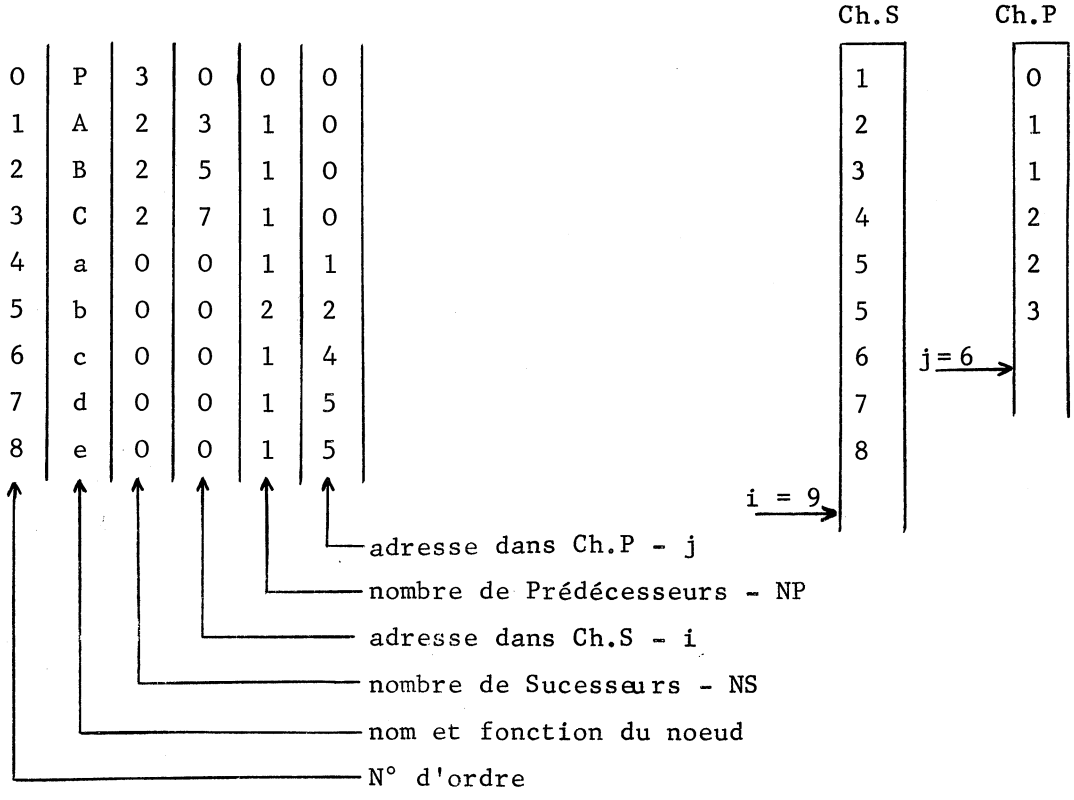
II - SUPPRESSION DE NOEUDS SPECIFIQUES A LA LANGUE SOURCE

Cette phase peut être inclue dans l'interprétation. Nous la décrirons ici en considérant qu'elle s'effectue en même temps que le changement de représentation de la structure qui est imposé par les considérations énumérées ci-dessus.

A - Structure ordonnée par niveau

Dans un graphe particulier considéré, le chemin entre la racine (qui est unique) et un noeud quelconque peut ne pas être unique. La "profondeur" du graphe est la longueur du plus long chemin entre la racine et une feuille. Il est possible que les feuilles ne soient pas toutes à la même distance de la racine. Nous appellerons "moment" du graphe le nombre de noeuds de celui-ci [8]. La matrice de connection, qui est une représentation commode du graphe n'est pas pratique dans notre cas, car elle nécessite des comparaisons en ligne ou en colonne selon la nature de la recherche (successeurs ou prédécesseurs).

Si nous adoptons la notion de niveaux, comptés à partir de la racine, il est possible d'organiser les noeuds par niveau. Si un noeud est atteint selon plusieurs chemins, son niveau est défini comme celui qui correspond au chemin minimum. De plus, deux chaînes auxiliaires décrivent les noeuds successeurs et prédécesseurs (chS et chP). Les tableaux suivants décrivent l'arbre donné au paragraphe I .



L'algorithme de transposition est simple : il faut balayer la liste en prenant le sucesseur le plus à gauche, puis successivement tous les noeuds de même niveau. La chaîne des noeuds, analogue à une chaîne postfixée, est construite de manière que chaque noeud n'apparaisse qu'une fois ; parallèlement, on fabrique Ch. S. Par contre, Ch. P. ne sera pas obtenue immédiatement car un noeud peut apparaître plusieurs fois dans Ch. S. Ainsi on crée au fur et à mesure une

chaîne, annexe à Ch. S, qui est réduite ensuite :

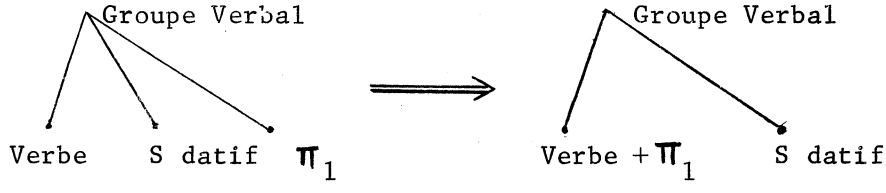
Ch. S	n° du prédécesseur
1	0
2	0
3	0
4	1
5	1
5	2
6	2
7	3
8	3

Les opérations suivantes sont exécutées successivement :

- Recherche des noeuds identiques dans Ch. S.
- Pour les noeuds ayant le même nombre de prédécesseurs, recherche des prédécesseurs identiques et rangement dans Ch. P avec indication de l'adresse j .

B - Suppression de noeuds

C'est, par exemple, le cas de la particule séparable en allemand ; en français, celle-ci forme un tout avec le verbe, en modifiant le sens de celui-ci ainsi que les valeurs des variables :



Ainsi, lorsque la feuille π_1 est supprimée, on connaîtra son prédécesseur grâce au procédé décrit ci-dessus ; l'information supplémentaire apportée par π_1 est alors redistribuée au successeur convenable, ou conservée partiellement au niveau du prédécesseur. Dans l'exemple précédent, la traduction sera différente lorsque

la particule est attachée au verbe, ce qui se traduit par un numéro de particule dans les variables du verbe.

Le composition peut s'appliquer aux structures figées en langue source (expressions ou locutions) qui forment un tout en langue cible. Les modifications provenant de la suppression d'éléments sont directement notées dans la chaîne d'éléments, Ch. S et Ch. P qui sont en construction.

III - CHOIX DES EQUIVALENTS

Nous avons vu (chapitre II, II - D) que le transfert de structure fait intervenir les particularités des mots ; nous avons besoin des codes syntaxiques associés à chacun. Schématiquement ayant une suite de mots, nous disposons aussi de la structure qui décrit les liaisons entre ces mots ; la première étape du transfert est donc une traduction "mot à mot" , les équivalents étant choisis selon les valeurs de variables sémantiques, syntaxiques et grammaticales.

A - Correspondance des unités lexicales

A une forme f correspond une forme f' déterminée en fonction du contexte de f (la notation ' distingue la langue cible). Nous avons repéré un ensemble de formes par son code classe et son numéro d' UL. Une partition a été introduite sur cet ensemble par les dérivations qui sont spécifiques au code classe. Un certain nombre de variables décrit f de manière que f' soit unique.

Dérivation $/_{CC}$ $\{\Delta_i\}$

Variables $\{v_j\} \longrightarrow$

- $\{vg_k\}$ selon catégorie syntaxique
- $\{vs_1\}$ selon emploi dans la phrase
- $\{v\sigma_h\}$ selon unité lexicale

La limite entre les dérivations et les variables n'est pas tranchée ; si la numérotation de Δ est reprise pour chaque code classe, on aura :

$$K_{u,v} = e_1 \ (CC, \Delta)$$

Une fois rappelée ces définitions, étant donné que nous n'avons pas un dictionnaire de formes, donc $f \longrightarrow f'$ est impossible, nous allons regarder ce que donne le transfert de la partition la plus grossière (au niveau des unités lexicales).

$$\begin{array}{ccc} \{f_i\} & \Longrightarrow & CC \ , \ UL \\ \vdots & & \downarrow \ \text{Transfert} \\ \{f'_i\} & \Longleftarrow & CC', \ UL' \end{array}$$

avec $\{\Delta_i\} \cup \{v_j\} \supseteq \{\Delta'_i\} \cup \{v'_j\}$

c'est à dire qu'il faut pouvoir choisir la forme adéquate dans l'ensemble f'_i .

Une redistribution des dérivations et des variables peut être faite, c'est à dire qu'une dérivation en langue source peut apparaître comme variable en langue cible (et inversement). Remarquons qu'une distinction nécessaire sur L_S peut être supprimée sur L_C (donc inclusion et non égalité).

Cette approche théorique n'est pas suffisante comme

nous allons le voir pour le transfert de la dérivation : lorsque la forme est une dérivation Δ de l'UL., Δ est conservé lors du transfert

$$<CC> <UL> <\Delta> \longrightarrow <CC'> <UL'> <\Delta>$$

EXEMPLES :

1 - Soit la forme allemande LEICHT appartenant à la classe adjectif ; avec la dérivation nulle, on obtient un adverbe. L'unité lexicale française correspond à la base FACILE (adjectif). On peut obtenir l'adverbe par la règle :

<forme de l'adverbe> ::= <base d'adjectif> <dés. fem.> MENT

qui s'applique dans ce cas et donne FACILEMENT.

2 - Prenons la forme FREUDLICH qui a la même dérivation que celle de l'exemple précédent. En français, nous traduirons par GENTIMENT qu'il n'est pas possible d'obtenir à partir de la base d'adjectif GENTIL selon la règle générale.

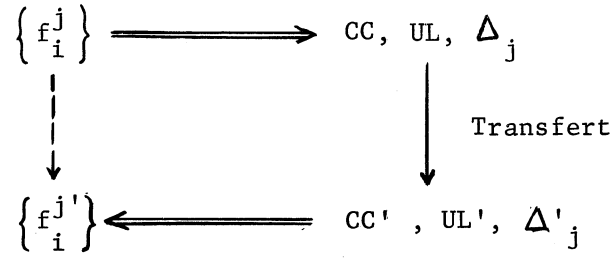
Si nous désirons conserver la correspondance bi-univoque entre les unités lexicales, la notion d'adverbe devra être notée en "variable".

3 - La dérivation Δ n'existe pas pour la classe considérée en français car il n'y a pas de règle assez générale.

Nous voyons que si ce principe de transfert était adopté, il n'y aurait de dérivations en langue source que si celles-ci existaient en langue cible.

B - Transfert de la dérivation

Au niveau de l'unité lexicale, les restrictions pour le transfert sont trop grandes, nous adopterons le tranfert sur le sous-ensemble de formes obtenu après partition par la dérivation



$'_j$ peut être nulle et la nouvelle catégorie syntaxique est donnée par $K'_{u,v} = e'_1 (CC' , \Delta'_j)$.

La condition de transfert est $\{v'_j\} \subseteq \{v_j\}$ qui permet de faire un choix entre les divers équivalents possibles.

C - Transfert selon les informations grammaticales et le sens

1 - Informations grammaticales

Nous donnerons une série d'exemples pour exposer ce problème.

- Verbe russe : pour certaines UL. la présence de CЯ , qui s'interprète comme une variable grammaticale notant le réfléchi, oblige à changer d'unité lexicale :

sans CЯ ⟶ DIRE

avec CЯ ⟶ S'EXPRIMER

- Verbe allemand : SICH (noté comme variable syntaxique) a le même effet que lors de l'exemple précédent :

HELFEN ⟶ AIDER

SICH HELFEN ⟶ SE TIRER D'AFFAIRE

- En allemand et en russe, selon le nombre on sera obligé de faire un choix ; ici par exemple en russe, on a :

au sing. ⟶ APPARTENANCE

au pluriel ⟶ { ACCESSOIRES / APPARTENANCES }

2 - Contexte

Le contexte détermine, dans certains cas, les valeurs de variables sémantiques qui permettent de choisir un équivalent parmi un groupe. Dans un premier temps, on pourra dire que le premier équivalent dans cette liste est "le plus probable" ; cette procédure sera utilisée avec l'interprétation. De toute manière pour réduire les équivalents, nous négligeons les synonymes.

3 - Transfert des variables

En théorie, le sens est l'invariant de la traduction

$$\{ v\sigma_k \} = \{ v\sigma'_k \}$$

Les valeurs des autres variables se transmettent, mais leur division en deux groupes peut être modifiée. Ainsi, le СЯ russe, obtenu lors de la morphologie, sera interprété en français comme une

variable syntaxique avec les valeurs : réfléchi, réciproque ou passif. Les choix entre les valeurs seront faits à l'aide des codes syntaxiques puis morphologiques. Si ces valeurs n'existent pas pour l'unité lexicale correspondante, il faut alors changer la tournure de la phrase à l'étape suivante, selon les règles de transfert de structure.

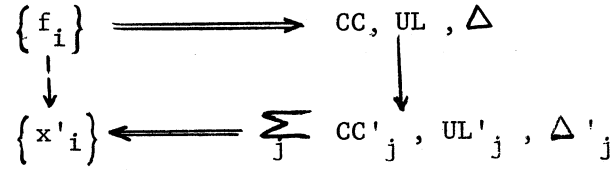
D - Dérivation spécifique à la langue cible

L'étude des langues sources ayant tiré partie des particularités propres à celles-ci sans s'occuper de la langue cible, pour cette dernière il a été fait de même.

Bien que le français soit une langue peu déclinée, un certain nombre de terminaisons ont été retenues ; du point de vue formel la dérivation signifie la concaténation d'un groupe de lettres à une base, mais nous avons préféré - comme en langue source - attacher un sens à la dérivation. Cette restriction peut être supprimée lorsque les essais montreront quelles sont exactement ses incidences et ses limites (voir chapitre V).

E - Locutions équivalentes

Plus généralement, une forme se traduit par une suite de formes à laquelle est appliquée une structure figée.



où un quelconque $x'_i = f'_1 \ldots\ldots f'_n$ avec la structure correspondante.

EXEMPLES

- Le verbe allemand SCHLEICHEN se traduit par AGIR SECRETEMENT. Les variables seront transférées sur le verbe AGIR (temps, mode, etc ...) mais il faut donner les informations se rapportant à SECRETEMENT ($K'_{u,v}$ d'adverbe, dérivation, etc ...) ainsi que la liaison syntaxique.

- Le mot composé ZWEIFLACHIG est traduit par A DEUX FACES. En fait, il est décomposé en deux éléments ZWEI et FLACHIG donnant respectivement DEUX et A .... FACES. Il faudra insérer entre les deux composants de la locution la "feuille" qui précède au même niveau.

Remarquons qu'il n'est pas rentable de définir une locution comme une unité lexicale, car les variables grammaticales ne se transmettent pas à tous les composants et nous aurions un dictionnaire trop volumineux.

F - Transfert automatique des dérivations

Les langues sources considérées admettent un certain nombre de dérivations automatiques (exemple : adjectif vers adverbe,..) ou non qui, dans la majorité des cas, obéissent à des règles générales en français.

1 - Dérivation

Nous avons la règle suivante :

$$CC, \Delta \longrightarrow CC' , \Delta'$$

Le numéro d'unité lexicale, en langue cible, est obtenu en faisant $\Delta = 0$

2 - Locution

Un certain nombre de dérivations automatiques en allemand se traduisent généralement à l'aide d'une tournure

$$CC\ ,\ UL\ ,\ \Delta \longrightarrow \sum_{j=1}^{n} CC'_j,\ UL'_j\ ,\ \Delta'_j$$

Parmi les éléments de la partie droite, il en existe un qui est obtenu par

$$CC\ ,\ UL\ ,\ \Delta_k \longrightarrow CC'_i\ ,\ UL'_i\ ,\ \Delta'_i$$

$$1 \leqslant i \leqslant n$$

Ainsi, on peut dire que Δ détermine deux exploitations :

- Déroulement de la chaîne équivalente
- $\Delta \longrightarrow \Delta_k$ et recherche de cet élément, celui-ci pouvant alors se ramener au cas 1 ci-dessus.

3 - Cas particulier de l'allemand : verbe à particule séparable

Le nombre des particules est assez élevé ; on peut espérer que l'adjonction des particules modifie d'une manière automatique la traduction du verbe seul. Il faut donc recenser les classes de particules qui ont une action identique. Si l'expérience montre que ces cas sont peu fréquents, la procédure générale de transfert décrite plus haut et fonction de l'unité lexicale sera utilisée.

IV - CONSULTATION DU DICTIONNAIRE D'EQUIVALENTS

A - Pseudo-texte d'entrée

Lors de l'analyse, une correspondance bi-univoque a été établie entre la numérotation lexicographique des bases et l'adresse d'une mémoire de la machine [9] .

<CC> <UL> <n° de base> ⟷ A

Le pseudo-texte décrit le graphe de structure et pour chaque feuille nous avons les renseignements lexicographiques et l'adresse A. Dans cette mémoire, sont venus se placer successivement le code morphologique, le code syntaxique puis le code interprétatif (ou éventuellement, le code sémantique). Cette adresse constitue aussi la racine du dictionnaire d'équivalents.

En langue cible, nous ne distinguerons pas entre les différentes bases d'une même unité lexicale, et d'une manière analogue nous associerons une adresse à chaque UL'

<N° identificateur '> ::= <CC'> <UL'>

<CC'> <UL'> ⟷ A'

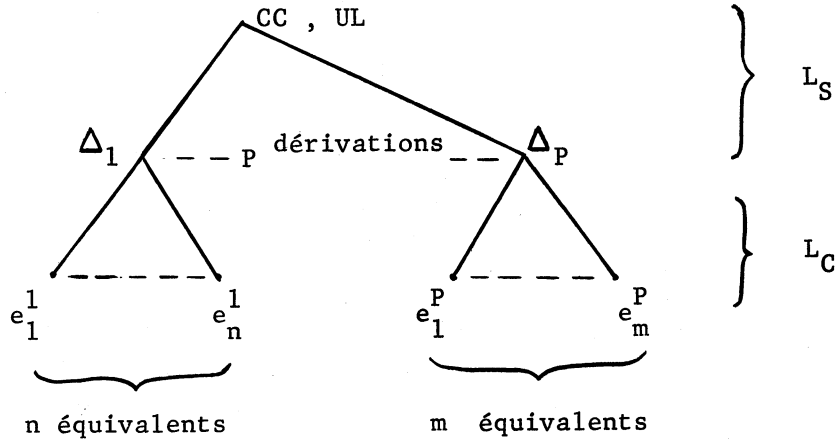
Remarquons qu'il est impossible de donner le numéro de base, qui dépend essentiellement des valeurs des variables, car nous ne disposons pas des codes morphologiques associés.

B - Organisation du dictionnaire d'équivalents

Dans un premier temps, nous n'étudierons pas le transfert des verbes à particules séparables.

1 - Arbre des équivalents

A chaque unité lexicale, appartenant à L_S , est associé l'arbre suivant :



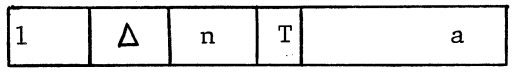
La dérivation calculée par l'analyse est le nom d'un noeud parmi les dérivations possibles sur l'unité lexicale. D'après les remarques du paragraphe précédent, les dérivations qui se transfèrent automatiquement ne seront pas notées, et seules les exceptions apparaîtront.

A partir de Δ_i , l'équivalent correct est choisi selon les informations grammaticales et les valeurs sémantiques. La présence de locutions doit être distinguée car elles nécessitent un traitement supplémentaire (insertion).

2 - Représentation machine du dictionnaire

Chaque noeud est figuré par une mémoire ; on distinguera entre les mémoires non terminales et les mémoires terminales. Chaque mémoire comporte trente six positions binaires (type ordinateur IBM 7090 - 7044).

Format des mémoires non terminales :



1 en position signe (-)

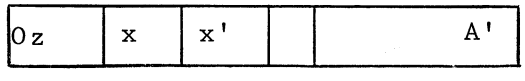
Δ dérivation en langue source numérotée par rapport à la classe morphologique

n nombre de mémoires dans la table des successeurs dont le début est en a

T numéro d'index pouvant faciliter la consultation de la table.

Si la mémoire A (racine) est de ce type, $\Delta \neq 0$ indique que les successeurs sont des dérivations possibles en langue source , $\Delta = 0$ que le choix entre les successeurs sera sémantique.

Format des mémoires terminales :



0 en position signe (+)

z en position 1

Premier cas : z = 0 il y a une unité lexicale corres-

x' = Δ' numéro de dérivation "transféré".

x = Δ s'il y a un seul équivalent

ou, lorsqu'il y a plusieurs équivalents les positions 2 - 11 sont utilisées pour noter les informations grammaticales (ou les variables sémantiques).

Deuxième cas : z = 1 traduction par une locution.

x' = c nombre de composants dans la locution.

x a la même signification que précédemment.

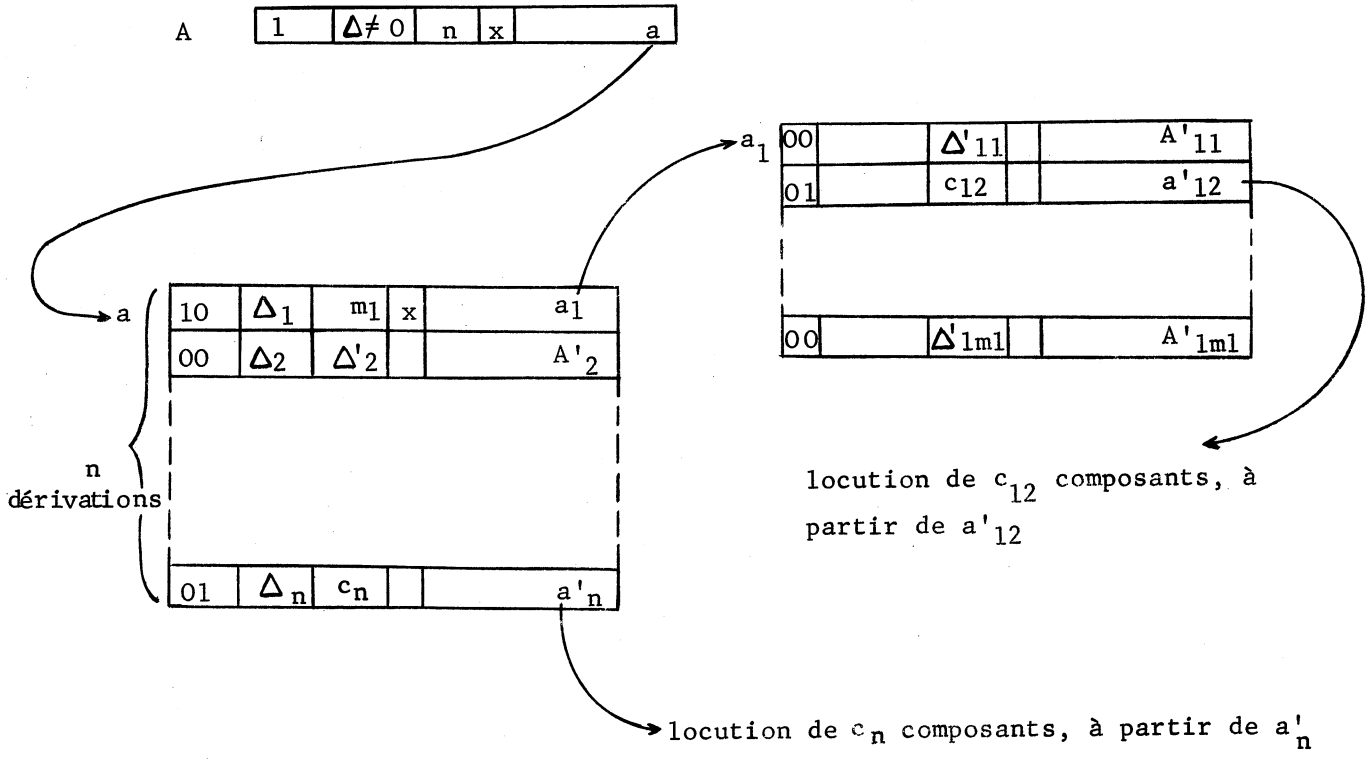
Profondeur de l'arbre des équivalents

Elle est toujours inférieure ou égale à 2 :

P = 0 un équivalent et pas d'exceptions pour le transfert automatique des dérivations.

P = 1 plusieurs équivalents sans dérivation ou plusieurs dérivations avec un seul équivalent.

P = 2 cas général qui est illustré par le schéma suivant :



Nous verrons plus loin la fabrication de ce dictionnaire à l'aide des feuilles d'équivalents données par les linguistes.

C - Exploitation du dictionnaire d'équivalents

1 - Encombrement

La consultation du dictionnaire se déroule en mémoire rapide du calculateur. Si le dictionnaire de langue source comporte 15000 bases, il y aura 15000 arbres des équivalents, non tous distincts. Les racines de ces arbres occupent 15000 mémoires, il faut placer aussi les noeuds de niveaux 1 et 2, ainsi que les locutions. Il n'est pas possible de dire actuellement si le dictionnaire, le programme et le noyau du système pourront tenir dans la mémoire rapide. Pratiquement, il faudra peut être couper les arbres au niveau 1 , un second passage permettant de terminer la consultation.

2 - Déroulement du programme

Le pseudo-texte d'entrée est traité en simultanéité. Il doit être mis sous la forme indiquée en paragraphe II. De plus, une transformation supplémentaire est nécessaire pour les numéros de dérivation : l'analyse morphologique note les dérivations à partir de la catégorie syntaxique et non de la classe morphologique.

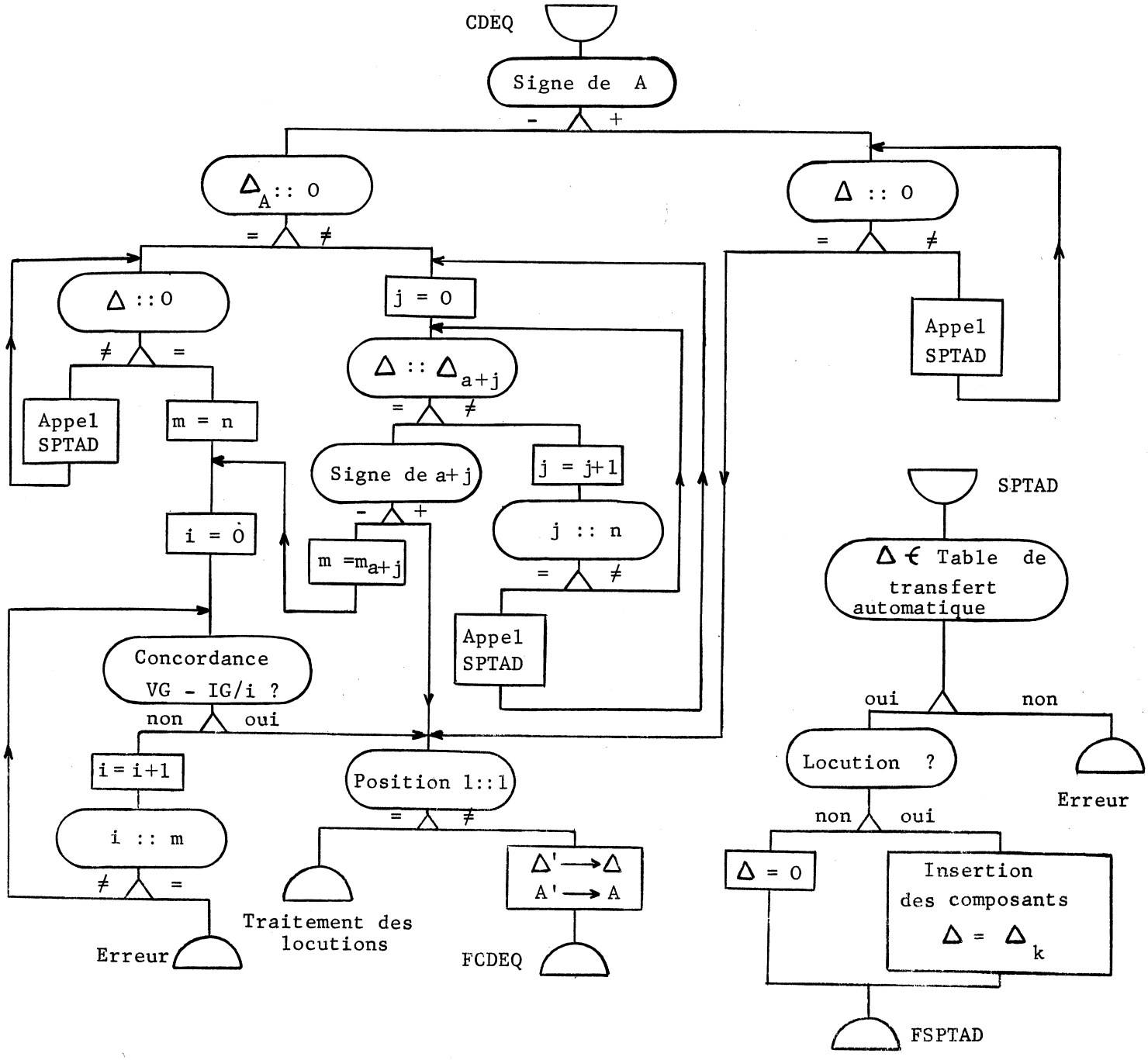
$$\Delta\,(K_{u,v}) = f(CC,\ K_{u,v})$$

Comme chaque feuille comporte les informations lexicographiques, CC et UL, nous pouvons rétablir les numéros de dérivation qui apparaissent dans le dictionnaire d'équivalents

$$g(\ \Delta\ (K_{u,v})\ ,\ K_{u,v}) = \Delta\ (CC)$$

L'exploitation du dictionnaire est effectuée selon l'organigramme de la page suivante. Le traitement des locutions est particulier car il faut insérer les composants, qui sont parfois "figés" dans la structure.

De plus, pour chaque élément du graphe de structure les variables sont redisposées dans un ordre donné qui est le même quelle que soit la langue origine.



CONSULTATION DICTIONNAIRE DES EQUIVALENTS - CDEQ - (exécuté pour chaque feuille).

Transfert automatique des dérivations - SPTAD - (appelé, dans certains cas, par CDEQ après balayage du dictionnaire, ce qui permet de résoudre les exceptions au transfert automatique ; par exemple, comparatifs irréguliers : meilleur, pire, etc ...).

D - Particularités de l'allemand

Nous avons noté le cas des verbes à particule séparable où la règle de transfert devient

$$CC_1 , UL_1 , \Delta_1 - CC_2 , UL_2 \longrightarrow \sum_j CC'_j , UL'_j , \Delta'_j$$

En simplifiant car CC_1 , CC_2 sont fixes

$$UL_1 , \Delta_1 - UL_2 \longrightarrow \sum_j CC'_j , UL'_j , \Delta'_j$$

En partie droite il y a, au moins une fois, la classe verbe. Les particules sont au nombre d'une centaine, environ.

Nous pourrons généraliser le dictionnaire de la manière suivante :

1 - Transfert automatique des particules

2 - Il existe des mémoires non terminales du type suivant :

11		Δ	p	T	P

p nombre de mémoires - donc de particules distinctes - dans la table P .

Si le dictionnaire n'est pas trop volumineux, cette recherche se fera en un seul passage ; dans le cas contraire nous pourrons prévoir deux passages.

V - DEPENDANCES SYNTAXIQUES

Nous abordons maintenant, l'étude des phases qui ne dépendent plus de la langue source.

A - Code syntaxique

Chaque "feuille" du pseudo-texte réfère à une adresse A' ; on trouvera dans cette mémoire le code syntaxique de l'unité lexicale correspondante. Etant donnée la structure interprétée construite sur des éléments "transférés" , nous choisirons la structure équivalente selon les particularités syntaxiques de ces éléments.

Le contenu du code syntaxique est déterminé par les linguistes ; par exemple, pour le verbe il contiendra des indications sur la transitivité, sur l'auxiliaire, etc ... Pour le substantif, nous aurons des renseignements sur le type de préposition admis selon le complément.

B - Règles

La liste des règles est établie par les linguistes d'après une étude des liaisons syntaxiques en langue cible. Les renseignements syntaxiques disponibles dans le pseudo-texte ont été fournis par l'analyse de la langue source. Nous aurons aussi les codes syntaxiques - qui sont donnés - et donc les codes de gouvernement et de dépendance qui conditionnent l'emploi des règles.

1 - Exploitation des valeurs des variables syntaxiques

Nous citerons quelques exemples de ce cas :

- Verbe à un temps composé :

V / TC, X ⟶ Auxiliaire /TS, X + Participe passé / Y

La correspondance TC - TS est donnée par la grammaire. L'auxiliaire est déterminé à l'aide du code syntaxique. Y est fonction de l'auxiliaire : genre et nombre du sujet si c'est le verbe ETRE , etc ...

- Verbe à la forme négative :

V / N, X ⟶ NE + V / P, X + PAS

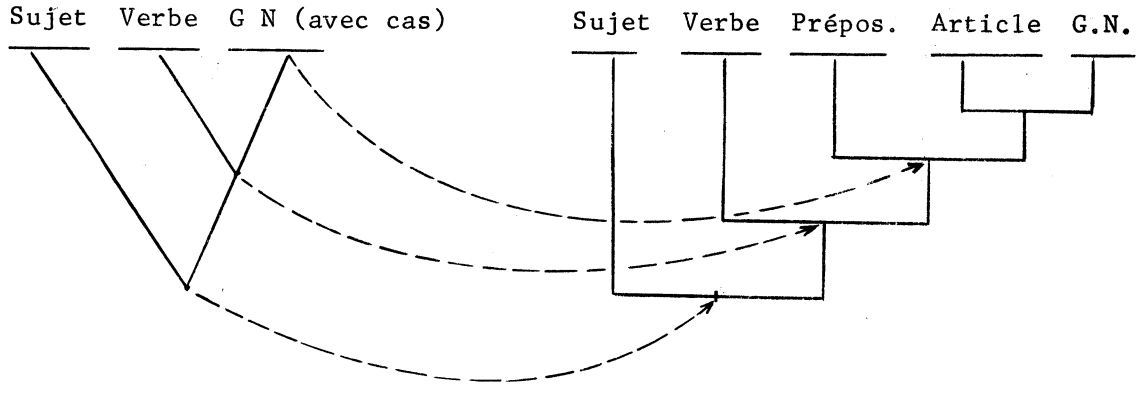
ou V / N, TC, X ⟶ NE + Auxiliaire / P, TS, X + PAS + Participe passé / Y

Ainsi, on ne conserve que les variables grammaticales qui modifient la forme des éléments terminaux, de manière à ce que la synthèse morphologique ne provoque pas d'insertion ou de suppression.

2 - Exploitation du code syntaxique

Il est évident que dans la phase 1, l'aide du code syntaxique est nécessaire. Nous voulons parler plus spécialement ici, de l'influence de ce code sur les éléments non terminaux de la structure. Par exemple, ayant reconnu une phrase à la voix passive, il faut la transformer en une phrase à la voix active si le verbe est intransitif. Ce qui aura pour conséquence de modifier les fonctions des noeuds ainsi que leur ordre.

3 - Eléments insérés



Un certain nombre d'éléments sont insérés lors de l'application des règles : auxiliaires, prépositions, articles, NE ... PAS , etc ... Ceux-ci appartiennent en majorité à des classes fermées ; les auxiliaires sont peu nombreux et doivent être distingués de l'ensemble des autres verbes. L'unité lexicale pourra être déterminée en fonction des valeurs des variables syntaxiques, on leur associera alors

$$g_1 (K'_{u,v} , \text{ v. s. }) \longrightarrow A'$$

Dans certains cas, il faut se limiter à la catégorie syntaxique

$$g_2 (K'_{u,v}) \longrightarrow A'$$

Le choix entre l'article défini et l'article indéfini, qui n'existent pas en Russe, a été examiné dans [19 - pages 11,12] Un certain nombre de règles ont été énoncées ; dans les cas ambigüs, où il faudrait analyser le sens du texte, on pourra choisir arbitrairement l'article indéfini.

VI - ACCORDS GRAMMATICAUX

A - Mots dominants

Les liaisons syntaxiques imposent l'accord des variables grammaticales, c'est à dire qu'un verbe est au même nombre que son sujet, etc ... Dans la phrase on distingue les mots dominants (ou gouverneurs) et les mots dominés (gouvernés). Du point de vue syntaxique, c'est le verbe qui conditionne la phrase (Tesnières). Par contre, lorsqu'il s'agit d'exploiter les liaisons syntaxiques existantes c'est le substantif qui est dominant, et plus généralement le sujet est prédominant. Les mots dominés sont l'adjectif, l'article, etc ...

Il faut par suite, reconnaître ces mots dominants et effectuer les accords qui apparaissent avec les mots dominés.

B - Modification des valeurs des variables grammaticales

Les valeurs des variables grammaticales ont été obtenues lors de l'analyse de la langue source ; on distingue :

- Des variables permanentes : leur valeur a été fournie par le dictionnaire.
- Des variables contingentes : Leur valeur a été calculée selon les désinences.

Cette distinction n'a pas de sens en langue cible, pour la synthèse, car on ne déduit pas les valeurs des variables de la composition d'une forme mais des indications de l'analyse et des conditions imposées par les particularités de la langue française. Pour préciser cela, prenons un exemple :

- Soit un substantif analysé en langue source ; nous aurons les renseignements suivants :

CC , UL

$K_{u,v}$ = Substantif commun , Δ = 0

Variable permanente Genre = Masculin

Variable contingente Nombre = Pluriel

Le transfert se fait selon la règle CC, UL $\longrightarrow$ CC' , UL'

Supposons que cette unité lexicale ne comporte qu'une base, dont la variable genre prend la valeur féminin uniquement. Nous serons amenés à modifier la valeur obtenue précédemment.

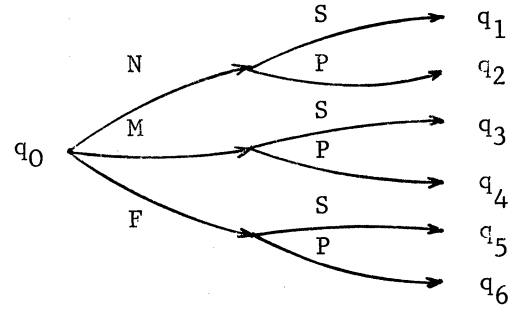
Le problème se présentera pour deux classes de mots :

- Les substantifs dont l'unité lexicale est défective
- Les verbes défectifs

Pour réaliser les accords grammaticaux il faut, en premier lieu, vérifier la validité des valeurs des variables grammaticales des mots dominants, puis en appliquant les liaisons syntaxiques compléter la détermination des mots dominés.

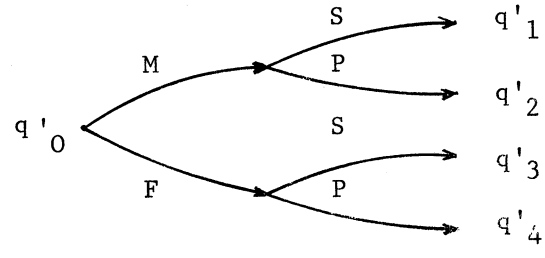
C - Genre et nombre du substantif

Pour les langues sources étudiées, nous allons considérer deux variables grammaticales du substantif, le Genre et le Nombre (G_S et N_S). Supposons que chaque variable ne puisse prendre qu'une valeur, c'est à dire qu'il n'y ait pas d'homographes. Lorsque ces variables prennent toutes les valeurs possibles, nous avons 6 formes :



(T)

Pour la langue cible, le genre et le nombre (G_C et N_C) admettent chacun deux valeurs, d'où quatre formes :



(T')

Si un ou plusieurs chemins sont interdits, l'unité lexicale est défective ; ainsi, on peut dire que chaque unité lexicale décrit un ensemble de formes qui est une combinaison quelconque des états terminaux.

Le neutre n'existant pas en langue cible, on peut considérer que le genre est indifférent et qu'il sera remplacé par le masculin s'il existe, ce qui revient à dire que le neutre est transformé de la même manière que le masculin. Un des chemins $q_o - q_i$ $(i > o)$ a été distingué par l'analyse ; le transfert a fait correspondre q_o à q'_o il s'agit de déterminer le chemin $q'_o - q'_j$ $(j > o)$ tel que $q_i \longrightarrow q'_j$.

Considérons le tableau suivant, où l'indice de colonne est donné par l'état terminal (attributs lexicaux) en langue source ;

Attributs lexicaux

Analyse / $f(L_C)$	MS q_1 , q_3	MP q_2 , q_4	FS q_5	FP q_6
0	0	0	0	0
1	0	0	0	MP
2	0	0	MS	0
3	0	0	MS	MP
4	0	FP	0	0
5	0	MS	0	FS
6	0	FP	MS	0
7	0	MS	MS	MS
8	FS	0	0	0
9	FS	0	0	MP
10	MP	0	FP	0
11	MP	0	MP	MP
12	FS	FP	0	0
13	FS	FS	0	FS
14	FP	FP	FP	0

l'indice de ligne est un numéro qui repère les combinaisons d'états terminaux en langue cible. L'index de ligne est calculé en affectant des poids à chaque forme :

"indice de ligne" = $2^3 \times \varphi(q'_1) + 2^2 \times \varphi(q'_2) + 2^1 \times \varphi(q'_3) + 2^0 \times \varphi(q'_4)$

avec $\varphi(q'_j) = 0$ si la forme correspondante existe, et $\varphi(q'_j) = 1$ dans le cas contraire. Si le composant du tableau trouvé est nul, les valeurs des variables ne changent pas, sinon leur nouvelle valeur est indiquée.

Pour réaliser cet ajustement des valeurs des variables nous disposons de l'indice de colonne (attributs lexicaux calculés) ; l'indice de ligne sera contenu dans le code syntaxique de l'U. L. ; c'est en effet une information attachée à l'unité lexicale car il rend compte des formes qui la composent.

Si nous avons plusieurs déterminations des attributs lexicaux (homographes internes), il faut consulter le tableau autant de fois qu'il y a d'homographes et comparer ensuite les nouvelles valeurs pour éliminer celles qui sont identiques. Il se peut que ayant deux homographes internes, la traduction conduise à la même forme.

D - Accords grammaticaux

Ayant déterminé de façon complète le mot dominant, il est possible de réaliser maintenant les accords en imposant les valeurs des variables à partir des liaisons syntaxiques existantes ; par exemple :

$$AD/\ X_1\ ,\ X_2\ +\ N/M,S \longrightarrow AD/M,S\ +\ N/M,S$$

Les verbes défectifs sont peu nombreux, en français, ils ne seront pas indexés lorsqu'il est possible de trouver un verbe synonyme, non défectif.

On peut remarquer que les verbes impersonnels (pleuvoir, falloir, ...) le sont aussi dans les langues considérées, ce résoud le problème.

Remarques :

- Si des homographes existent, chacun est placé dans le pseudo-texte de sortie avec une valeur unique des variables grammaticales.

- Les études linguistiques pour l'établissement du code syntaxique font apparaître l'existence de "noyaux" dans la phrase ; comme exemples, citons le groupe nominal, le groupe verbal, etc ... Les liaisons entre ces noyaux sont données par des règles syntaxiques nécessitant des accords grammaticaux.

E - Pseudo-texte de sortie

Au début du transfert de structure nous disposions d'une structure portant sur des "feuilles". Au cours de cette étape nous avons transformé cette structure et nous avons exploité les incidences grammaticales qu'elle comporte, de manière à reporter toute l'information sur les feuilles.

Ainsi, l'ordre des noeuds a été modifié, leurs caractéristiques ont été adaptées dans le but de la synthèse. Nous disposons maintenant d'un "syntagme élémentaire" analogue à celui obtenu lors de l'analyse qui a, en plus, profité des informations syntaxiques.

On peut donc considérer que le pseudo-texte de sortie est formé uniquement des éléments terminaux du graphe, apparaissant dans l'ordre normal où seront générés les mots du texte associé.

CHAPITRE IV

CONSULTATION DU DICTIONNAIRE MORPHOLOGIQUE

I - CONTENU DU DICTIONNAIRE

Le dictionnaire contient tous les renseignements morphologiques et lexicographiques nécessaires à la synthèse des formes ; en particulier, on y trouve les suites de lettres décrivant les bases indexées par les linguistes.

A - Eléments linguistiques du dictionnaire

Ils apparaissent sur les feuilles d'indexage des linguistes.

1 - Découpage

Dans le but de réduire le volume du dictionnaire les formes sont découpées selon les types de composants suivants [23] :

- Base B
- Préfixe P
- Suffixe S
- Désinence δ
- Particule λ

Les formes peuvent être de trois types :

- λ correspond aux ponctuations , ...
- P B δ_1 S δ_2 δ_3 décrit les mots simples de la langue.

DES HABILL AGE S
P B S δ_3

- B' δ'_1 δ'_2 λ' B δ_1 δ_2 est valable pour les mots composés (substantifs et adjectifs).

SOURD E S - MUET TE S
B' δ'_1 δ'_2 λ' B δ_1 δ_2

(λ' ne peut prendre que les valeurs " - " ou " # DE # ")

Lorsque le système de désinences n'est pas assez riche, les deux bases sont indexées :

Exemple : CORAIL
CORAUX

Nous verrons que la distinction d'emploi de ces bases est notée dans le code morphologique.

2 - Code morphologique

Le code morphologique donne les caractéristiques de la base. Il indique les systèmes de désinence qui doivent être appliqués selon les valeurs des variables grammaticales et précise les possibilités d'utilisation de la base associée ; par exemple, le code morphologique de CORAIL donne la valeur MASCULIN pour la variable GENRE et SING pour la variable NOMBRE.

La liste des variables grammaticales associée à chaque classe terminale est donnée dans [23] .

3 - Remarques sur la dérivation

- Une dérivation donnée impose un nouveau code morphologique, qui est implicite dans la dérivation. Ainsi, il est faux de dire que dans les deux cas suivants la dérivation est la même :

CHANT	EUR	(1)
B	S	
CLAM	EUR	(2)
B	S	

car la forme (1) est masculine, (2) étant féminine.

- Nous avons vu le transfert de la dérivation au chapitre précédent. Au niveau morphologique il n'y a pas vérification, c'est à dire que si une dérivation est donnée, elle est appliquée automatiquement à l'unité lexicale correspondante. Pour cela, la feuille d'indexage comporte une partie code dérivation. Un certain nombre de restrictions ont été imposées :

- Les noms composés n'ont pas de code dérivation

- Les verbes à bases multiples n'admettent qu'une partie des dérivations applicables aux verbes à base unique.

B - Article de dictionnaire morphologique

A chaque unité lexicale est associé un article de dictionnaire

<article de dict. > ::= <en-tête d'article> <corps d'article>
<en-tête d'article> ::= <l > <CC' > <UL'>
<corps d'article > ::= $<B_1>$ $<c\mu_1>$ ..... $<B_n>$ $<c\mu_n>$

L'en-tête a un format fixe ; l est la longueur du corps d'article, en nombre entier de mémoires. Si l'unité lexicale comprend n bases, le corps d'article contient la première base, son code morphologique, puis la deuxième base, son code morphologique et ainsi de suite.

L'en-tête sert à repérer le corps d'article qui seul est utile pour la synthèse morphologique. Le dictionnaire est sur bande magnétique, chaque enregistrement comprenant plusieurs articles. Nous verrons plus loin le format du corps d'article.

II - PROCEDES DE CONSULTATION D'UN DICTIONNAIRE SUR BANDE

Ces procédés ont été exposés en détail dans [9] . Nous allons les rappeler brièvement de manière à choisir celui qui présente le plus d'avantages pour le dictionnaire décrit ci-dessus.

A - Dictionnaire en mémoire rapide

Le dictionnaire est mis entièrement en mémoire rapide. Pour retrouver les articles utiles, il est nécessaire de conserver l'en-tête. Pour un dictionnaire de 15 000 bases le volume occupé dépasse largement la capacité du calculateur. De plus, la recherche des articles est faite par comparaisons successives qui allongent le temps d'exécution. Le procédé adopté lors de l'analyse utilise une disposition en arbres qui évite les répétitions, et il n'était pas nécessaire d'avoir tous les renseignements morphologiques. Pour ces diverses raisons, il ne peut s'appliquer ici.

B - Dictionnaire sur bande

Nous supposerons que les articles ont été triés par code classe, puis par numéro d'unité lexicale.

1 - Lecture et synthèse morphologique

Le pseudo-texte étant lu, nous avons besoin d'un article défini par <CC' , UL'> . La bande est déroulée jusqu'à ce qu'on ait trouvé l'article correspondant. Remarquons que pour effectuer la synthèse d'une forme, il faut disposer de plusieurs articles de dictionnaire car on peut avoir élision ou contraction, d'où nécessité de procéder à un certain stockage. L'inconvénient majeur de cette méthode est que la bande magnétique est sollicitée très souvent, au hasard, dans un sens ou dans l'autre ; les nombreux espacements ainsi imposés allongent considérablement le temps de consultation.

2 - Lecture globale

On lit d'abord tous les articles demandés par le pseudo-texte, la synthèse s'effectuant seulement lorsque cette lecture a

été terminée . Ceci nécessite un passage supplémentaire du pseudo-texte mais permet l'exécution du traitement final en mémoire rapide. Cependant, si la même unité lexicale apparait plusieurs fois dans le pseudo-texte, l'article correspondant apparaîtra autant de fois en machine ; de plus les mots du texte étant dans un ordre aléatoire, le même inconvénient que précédemment existe.

C - Dictionnaire sélectif

C'est la méthode adoptée qui est dérivée du processus décrit par Lamb [3] . Elle supprime les inconvénients notés plus haut ; elle permet une consultation séquentielle du dictionnaire et un traitement en mémoire rapide avec une compression maximum du dictionnaire morphologique.

III - LISTE INTERMEDIAIRE

A - Objectifs

Nous rappelons ci-dessous les objectifs à atteindre pour accélérer la consultation du dictionnaire et la synthèse morphologique :

1 - Lecture séquentielle de la bande dictionnaire.
2 - Dictionnaire "utile" en mémoire rapide du calculateur.
3 - Chaque corps d'article apparait une seule fois.

Les conditions 2 et 3 impliquent un encombrement minimum de la mémoire rapide. Pour réaliser ces impératifs nous utiliserons la technique de la"liste intermédiaire" qui sera employée conjointement

avec le pseudo-texte en premier lieu, puis servira à choisir dans le dictionnaire les articles intéressants.

B - Contenu de la liste intermédiaire

Au chapitre précédent (chapitre III - IV-A) nous avons associé une mémoire d'adresse A' à chaque numéro d'unité lexicale

$$CC' , UL' \longleftrightarrow A'$$

Pour deux unités lexicales différentes si $CC'_1, UL'_1 < CC'_2 , UL'_2$ alors on doit avoir $A'_1 < A'_2$. Ceci signifie que la liste intermédiaire est triée par code classe puis par numéro UL' ; si on dispose de micro-glossaires spécialisés, on associera à chacun une liste intermédiaire, disposée dans le même ordre que le micro-glossaire ; en particulier, il n'est pas nécessaire que les unités lexicales soient numérotées en séquence.

Le contenu de la mémoire A' est le suivant

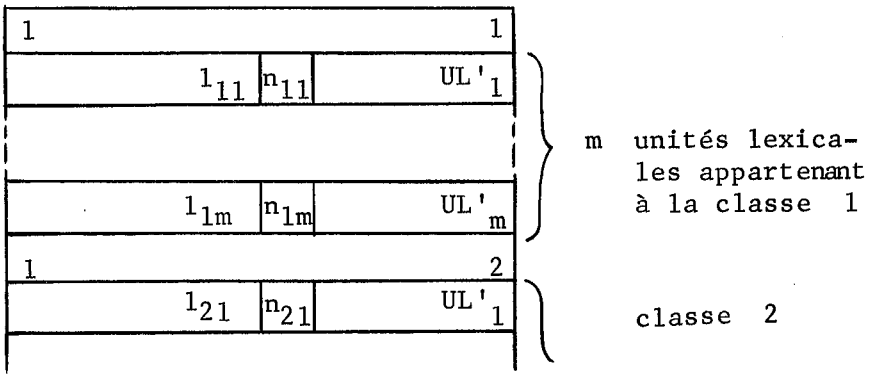
0	1	n	n° UL'

- La position signe est nulle
- 1 est la longueur du corps d'article correspondant (identique dans l'en-tête d'article)
- n indique le nombre de bases qui se trouvent dans le corps d'article donc qui appartiennent à la même unité lexicale.

Les classes morphologiques sont séparées par un marquant ; c'est une mémoire dont la position signe est à 1 , l'adresse

contenant la valeur du nouveau code classe.

L'ensemble de telles mémoires des deux types ci-dessus, constitue la "liste intermédiaire" par analogie avec l'appellation donnée par Lamb 3 . Schématiquement cette liste se présentera sous la forme :



Remarques :

Pour certaines classes fermées (exemple : pronoms, articles, etc ...) le nombre de bases distinctes de l'unité lexicale est donné - implicitement - par la catégorie syntaxique, qui est, pour des raisons d'indexage, distinguée aussi par une combinaison particulière de CC', UL'.. Dans ce cas, n_{ij} prendra la valeur arbitraire 1 .

C - Exploitation de la liste intermédiaire

On lit le pseudo-texte d'entrée dans lequel nous nous intéressons seulement aux adresses A' qui ont été fournies par le dictionnaire d'équivalents. Si la mémoire correspondante est de signe "plus" , l'unité lexicale n'a pas encore été réclamée par la partie antérieure du pseudo-texte. Cette position signe est forcée à 1 et une "adresse temporaire" (AT) est calculée de la manière suivante :

AT := EDS

EDS := EDS + 1

EDS est la valeur de l'encombrement actuel du dictionnaire déjà sélectionné ; AT vient remplacer 1 dans la mémoire A' et, dans le pseudo-texte se met en lieu et place de A' . De plus, n et le couple CC', UL' sont insérés dans le pseudo-texte de sortie. Le calcul de CC' est effectué par comparaison dans une table auxiliaire :

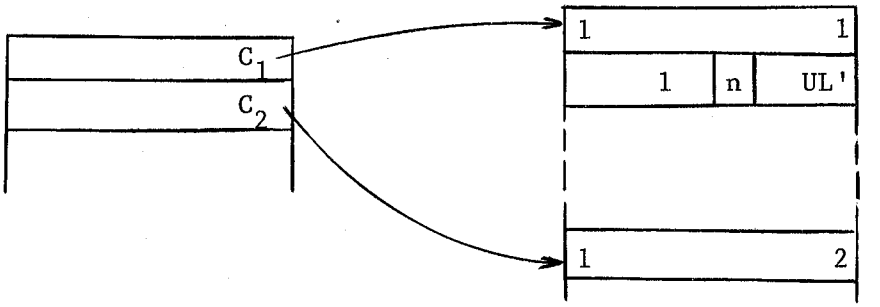


Table auxiliaire Liste intermédiaire

on détermine $C_i < A'$ tel que $C_{i+1} > A'$

Si la position signe de la mémoire A' est "moins" , l'adresse temporaire n'est pas recalculée.

On a vu qu'il pouvait subsister encore des ambiguïtés sur le choix de l'unité lexicale (chapitre II) qui pourront peut-être se résoudre en morphologie. Pour cette raison, quelques classes fermées sont sélectionnées systématiquement, c'est à dire que toutes les mémoires correspondantes ont leur signe forcé à 1, et on trouve directement l'adresse temporaire. De cette manière, l'encombrement à l'origine EDS_0 n'est pas nul mais fixé à une valeur donnée.

Il est vérifié à chaque nouveau calcul de l'encombrement que celui-ci ne dépasse pas la capacité de mémoire rapide qui lui est imposée par la taille du programme et les réservations nécessitées

par le système de programmation. Dans un tel cas, il est nécessaire d'effectuer le passage à la phase de génération et de prévoir une procédure de reprise de la traduction à l'endroit où l'on s'est arrêté. L'organigramme détaillé est donné ci-après. La liste intermédiaire pour un dictionnaire de 12 000 unités lexicales occupe 12 000 mémoires du calculateur ; l'exploitation peut donc être faite en un seul passage .

D - Compression de la liste intermédiaire

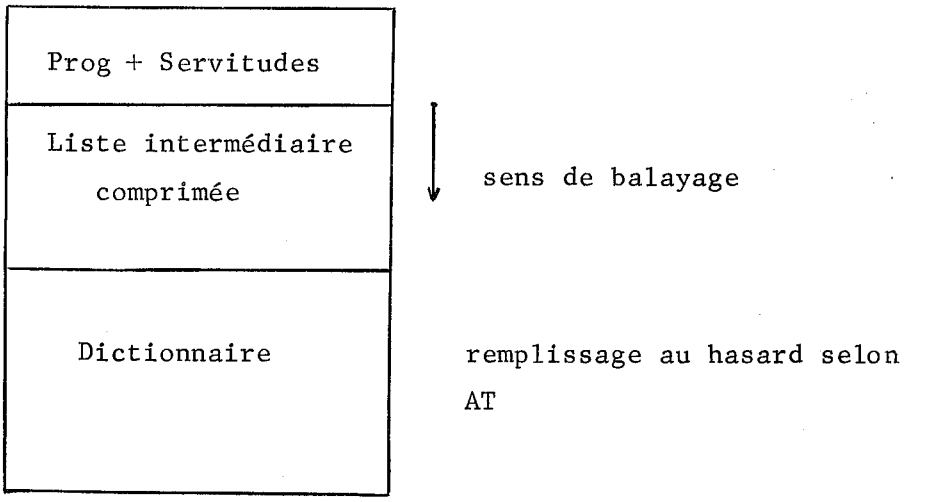
Les articles à sélectionner sont repérés par leur numéro d'unité lexicale dont la mémoire associée est signée "moins" . La liste intermédiaire contient de nombreux "blancs" qui sont inutiles. On effectue le rassemblement des mémoires (-) dans une partie de la mémoire rapide, ce qui libère le reste des mémoires pour le chargement ultérieur du dictionnaire.

IV - SELECTION DES ARTICLES DU DICTIONNAIRE MORPHOLOGIQUE

Cette phase ne fait pas appel au pseudo-texte dont le support d'information peut être repositionné durant le traitement.

La liste intermédiaire "comprimée" permet de choisir les articles adéquats selon leur en-tête. Le corps d'article est alors placé à partir de l'adresse temporaire calculée précédemment. La liste intermédiaire et le dictionnaire morphologique ayant été triés selon les mêmes critères (CC, UL), le déroulement de la bande est séquentiel.

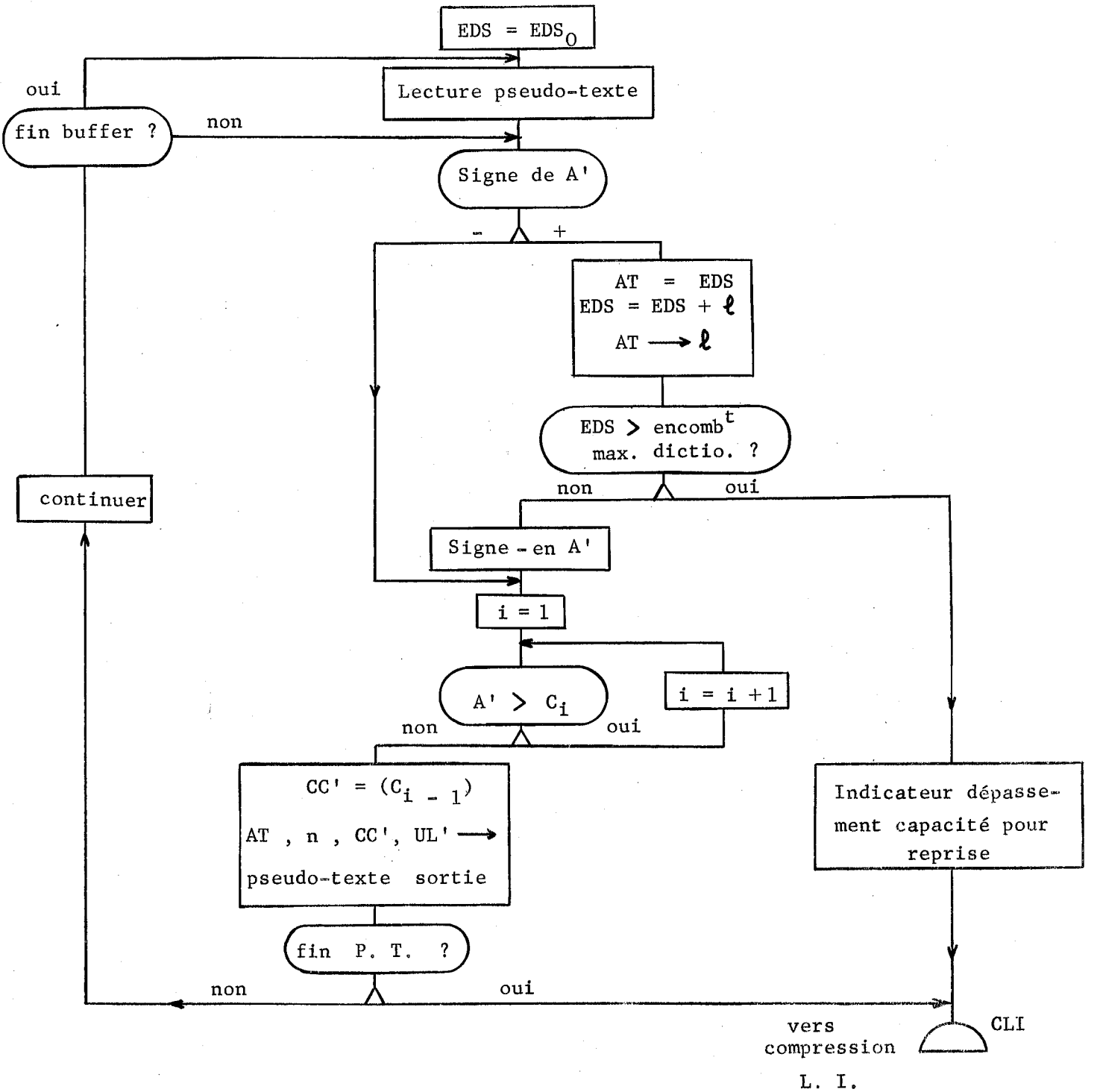
Nous montrons ci-dessous le principe d'occupation des mémoires pendant la sélection :



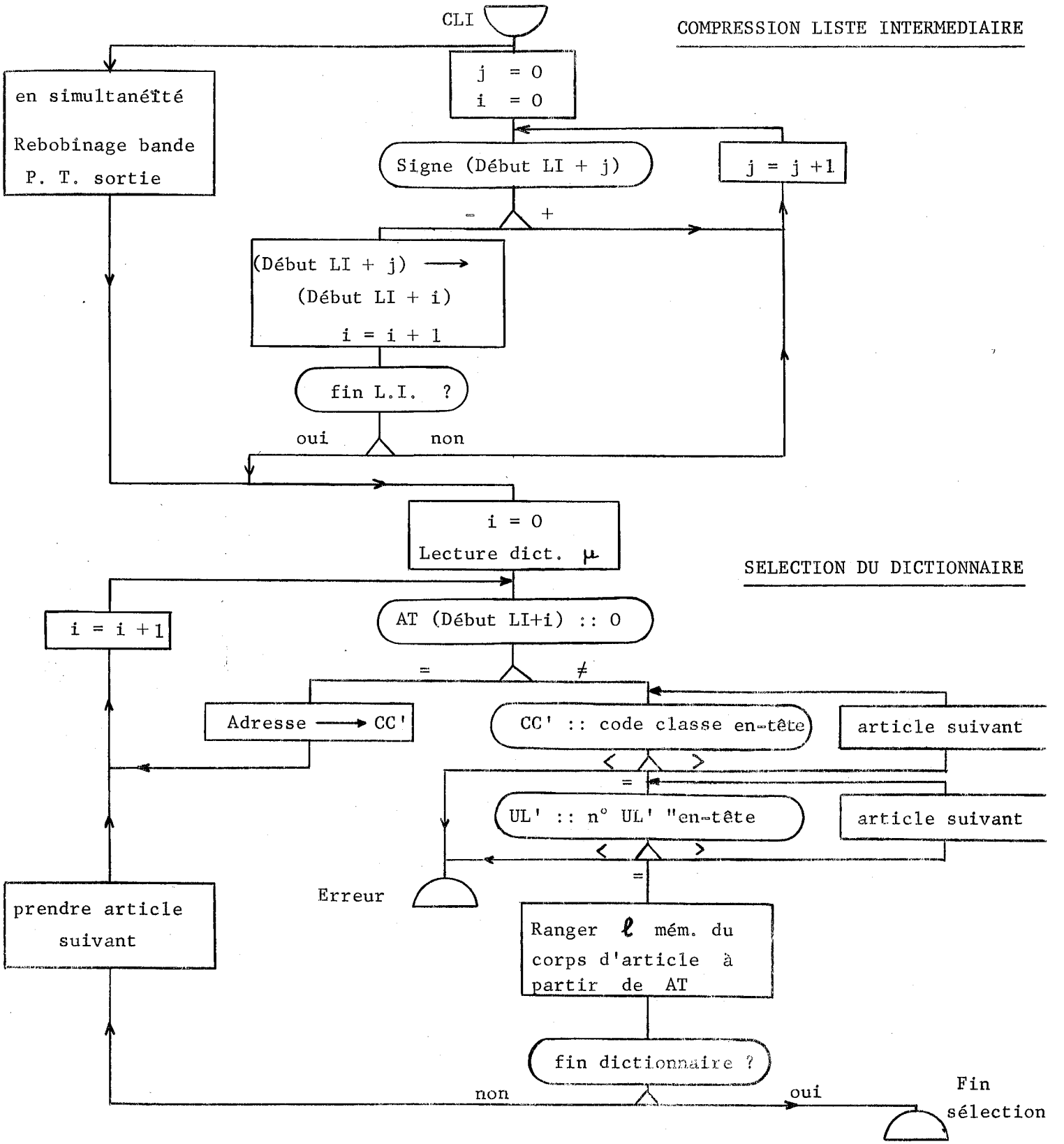
Les programmes décrits, dont les organigrammes sont en page suivante, ont été mis au point sur ordinateur IBM 7090 avec un dictionnaire réduit et sans entrée - sortie sur bandes magnétiques.

Intérêt de ce procédé

Selon Lamb, on peut estimer que la traduction d'un texte nécessite un nombre de bases égal à 10 % , environ, du nombre d'occurrences. Pour 50000 occurrences d'entrée, le nombre d'unités lexicales différentes sera à peu près de 4500. Si on considère qu'en moyenne le corps d'article occupe cinq mémoires, le dictionnaire "utile" tiendra facilement en mémoire rapide. Nous verrons, au chapitre suivant un procédé pour diminuer l'encombrement de chaque article.



EXPLOITATION DE LA LISTE INTERMEDIAIRE


CLI
COMPRESSION LISTE INTERMEDIAIRE
en simultanéité
Rebobinage bande
P. T. sortie
j = 0
i = 0
Signe (Début LI + j)
j = j +1
-
+
(Début LI + j) ⟶
(Début LI + i)
i = i + 1
fin L.I. ?
oui
non
i = 0
Lecture dict. μ
SELECTION DU DICTIONNAIRE
i = i +1
AT (Début LI+i) :: 0
=
≠
Adresse ⟶ CC'
CC' :: code classe en-tête
article suivant
=
UL' :: n° UL' "en-tête
article suivant
=
Erreur
prendre article
suivant
Ranger ℓ mém. du
corps d'article à
partir de AT
fin dictionnaire ?
non
oui
Fin
sélection

CHAPITRE V

S Y N T H E S E M O R P H O L O G I Q U E

I - FORME MACHINE DU CORPS D'ARTICLE

Pour réduire l'encombrement total du dictionnaire dans le but de traiter des textes assez volumineux, nous avons - au chapitre précédent - montré comment nous supprimions les articles superflus. Cette optimisation étant faite, nous allons porter notre attention sur l'encombrement du corps d'article de dictionnaire morphologique et essayer de réduire ses dimensions.

A - Compression des bases

Le problème étant différent, il est impossible de procéder à un découpage global des bases en arbres et en listes [9] .

Cependant, si une unité lexicale comporte plusieurs bases, en général ces bases débutent par des suites de lettres identiques ; pour les bases CORAIL et CORAUX on peut écrire :

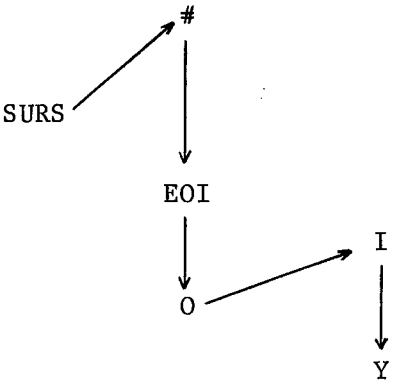
Pour les verbes irréguliers, la comparaison est encore plus flagrante :

Exemple :

L'unité lexicale correspondant au verbe SURSEOIR a les quatre bases suivantes :

S U R S|
S U R S|E O I
S U R S|O|I
S U R S|O|Y

d'où la décomposition :

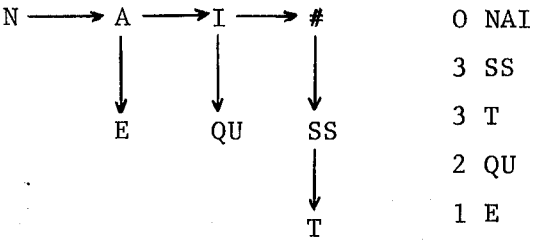


Cette représentation se traduit sous forme séquentielle en associant à chaque point terminal, le nombre de lettres identiques dans la succession des bases.

0 S U R S
4 E O I
4 O I
5 Y

Nous donnons comme exemple, la représentation des bases

du verbe NAITRE :



B - Code morphologique

A chaque base est associé un code morphologique ; il suivra immédiatement la base, c'est à dire qu'il n'a pas de position fixe en machine. Le code morphologique machine occupe trois caractères (c'est à dire dix huit positions binaires), quelle que soit la classe morphologique de l'unité lexicale correspondante (voir Annexe).

C - Détail du corps d'article

Les différentes bases de l'unité lexicale sont classées par ordre alphabétique ; en plus, de la suite de lettres une base comporte divers renseignements :

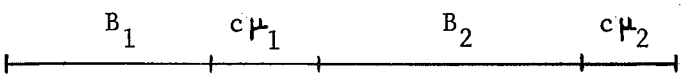
$$B_i \longrightarrow \pm K \underbrace{XXX \ldots . X}_{K \text{ caractères}} < 32$$

Signe + ⟶ base non "comprimée"

Signe - ⟶ base "comprimée" : le premier X est numérique, il indique le nombre de lettres identiques dans la base précédente. Pour la première base, le signe est toujours + .

Le corps d'article se présentera donc, par exemple, sous la forme :

$$+ k_1 \underbrace{X\ X\ X\ X}_{k_1}\ \underbrace{Y\ Y\ Y}_{3} \quad - k_2\ h_1 \underbrace{X\ X\ X\ X}_{k_2}\ \underbrace{Y\ Y\ Y}_{3} \quad \ldots$$



La longueur, en nombre de caractères, du corps d'article pour n bases est :

$$L = n + \sum_{i=1}^{n} k_i + 3\ n$$

et l , noté dans l'en-tête et la liste intermédiaire, est donné par

$$l = \text{entier immédiatement supérieur à } \frac{L}{6}$$

L'économie ainsi réalisé par la compression des bases est de l'ordre de 15 % pour les unités lexicales à plusieurs bases. Pour un calculateur dont l'unité de traitement serait le caractère, il serait intéressant de donner L et non l , c'est à dire que les corps d'article seraient placés les uns à la suite des autres. Avec le calculateur dont nous disposons la complexité du problème supprimerait ces avantages car la sélection du dictionnaire prendrait un temps plus long. Remarquons, que ici le temps nécessaire à la recomposition des formes est légèrement plus grand.

II - SYNTHESE MORPHOLOGIQUE [23]

A - Principe

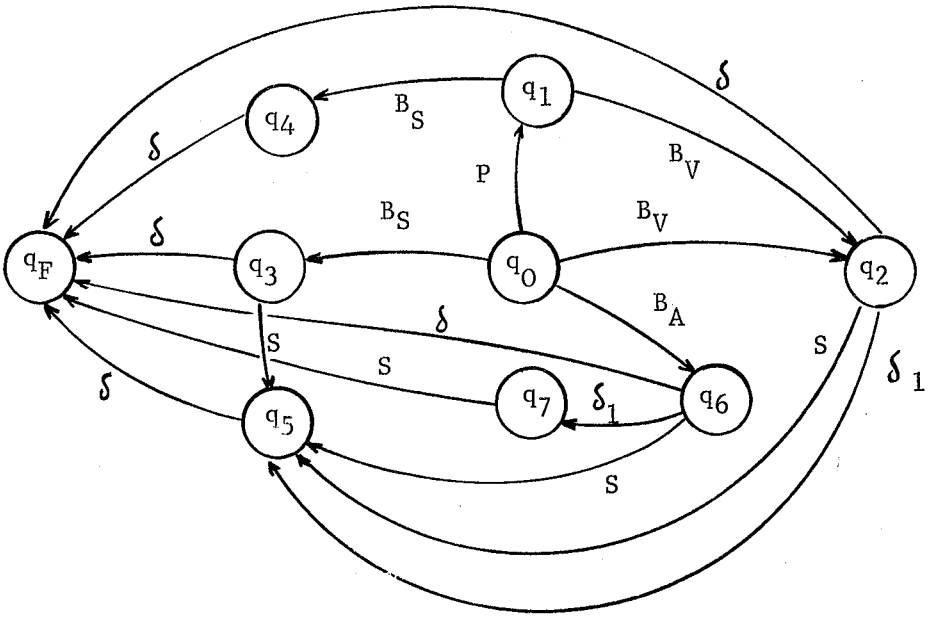
Le pseudo-texte est constitué par une chaîne d'éléments E_i. A chaque E_i est associé un article du dictionnaire morphologique. En fonction des valeurs de certaines variables grammaticales de E_i et du code morphologique associé à la base, nous choisirons la base adéquate pour la synthèse. Ensuite, il sera possible de trouver les désinences selon les mêmes critères. Il n'y a pas de test à faire sur la dérivation, celle-ci étant toujours possible lorsqu'elle est demandée (dictionnaire d'équivalents). Rappelons que l'adresse du corps d'article apparait dans le pseudo-texte : elle a été calculée au moment de l'exploitation de la liste intermédiaire.

B - Grammaire d'états finis

La description des formes, à l'exception des mots composés de la langue, est donnée par la grammaire suivante :

- On distingue B_A : base d'adjectif
  B_S : base de substantif
  B_V : base de verbe

- D'autre part $\delta = \delta_2 \ \delta_3$



L'état initial est l'état q_0 , l'état final q_F ; Nous n'avons pas distingué les désinences, cette grammaire pouvant être raffinée selon les différentes classes morphologiques.

Le détail des classes morphologiques, des catégories syntaxiques, des noms et des valeurs des variables est donné dans un document CETA, analogue à ceux qui ont été établis pour les langues sources analysées.

III - ELISION, CONTRACTION ET PROBLEMES CONNEXES

A - Règles

1 - Elision

Les règles se mettent sous la forme générale

ab ⟶ cb

a, b et c étant des chaînes construites sur le vocabulaire terminal.

Ce sont des règles phonétiques, car la dernière lettre de a et la première lettre de b sont des voyelles. On peut répertorier les classes auxquelles appartient la chaîne a qui subit l'élision, et dans une certaine mesure les chaînes b qui provoquent l'élision.

Nous avons introduit la notion de code élision pour les chaînes de type a, et cette indication est portée dans le code morphologique. Lorsqu'une telle chaîne doit être générée, il faut la mettre en réserve et attendre le mot suivant. Les lettres qui provoquent l'élision ont une représentation machine distincte ; nous les noterons par * .

Exemples de règles d'élision

a/ Article défini, E + * b ⟶ l' b
a/ Pron. personnel, E + * b ⟶ c' b

Exceptions :

- Dans certains cas, dépendant du verbe, les pronons JE, LE et LA ne subissent pas l'élision :

  - Lorsque JE suit un verbe à un temps simple
  - Lorsque LE et LA suivent un verbe à l'impératif et sont compléments de ce verbe.

La règle correspondante est donc "saturée" par le verbe et libérée au mot suivant, dans le premier cas ; d'autre part les pronoms LE et LA "saturent" cette règle quand la variable CAS a la valeur accusatif tonique, cette valeur étant donnée par une règle de

syntaxe :

V / Imp + PP / 3e pers, Acc ⟶ V/ Imp + PP / 3e pers, Acc T.

- Pour la conjonction SI, on a une règle particulière ne s'appliquant que pour un pronon personnel, puisque dans ce cas les exceptions sont tous les autres mots.

2 - Contraction

On peut écrire les règles de la manière suivante :

ab ⟶ c

où a est une préposition : A ou DE

b est de la forme LE x x x ou LES x x x x ; b appartient à des classes fermées (article, pronom). Un code contraction a aussi été créé pour ces classes dans le code morphologique.

Remarque

L'élision précède la contraction, ce qui veut dire que dans le cas de l'article défini LE il faut attendre que le mot suivant soit généré pour appliquer la règle ; il y a lieu de stocker les renseignements.

3 - Problèmes connexes

- Modification de certains adjectifs, au singulier, devant un substantif

$a/ Adj, S + * b/ Subst. ⟶ c + b

- Les valeurs des variables grammaticales des adjectifs possessifs

(genre, nombre du possesseur, nombre du possédé) suivent la règle

a/ Adj possessif, F,S,S, + * b ⟶ a/Adj possessif,M,S,S + b

- Pour l'adjectif démonstratif

CE + * b ⟶ CET + b

B - Principe de réalisation

Comme nous l'avons vu, le code morphologique, pour certaines classes fermées, contient des indications qui autorisent l'application ou non de ces règles. De plus, il faut prévoir le rangement temporaire de l'élément de gauche de la partie gauche, parfois même de l'élément de droite. L'application de la règle est déclenchée par l'apparition de l'élément de droite de la partie gauche (élision, problèmes connexes) ou par le mot suivant (contraction).

Nous effectuerons le rangement dans une pile puisque c'est le dernier élément rentré qui sera modifié, par contre la "libération" de cette pile aura lieu à partir du premier rentré. Les cas d'exception empêchent le stockage intermédiaire. La réalisation pratique d'un tel procédé sera décrite au chapitre suivant.

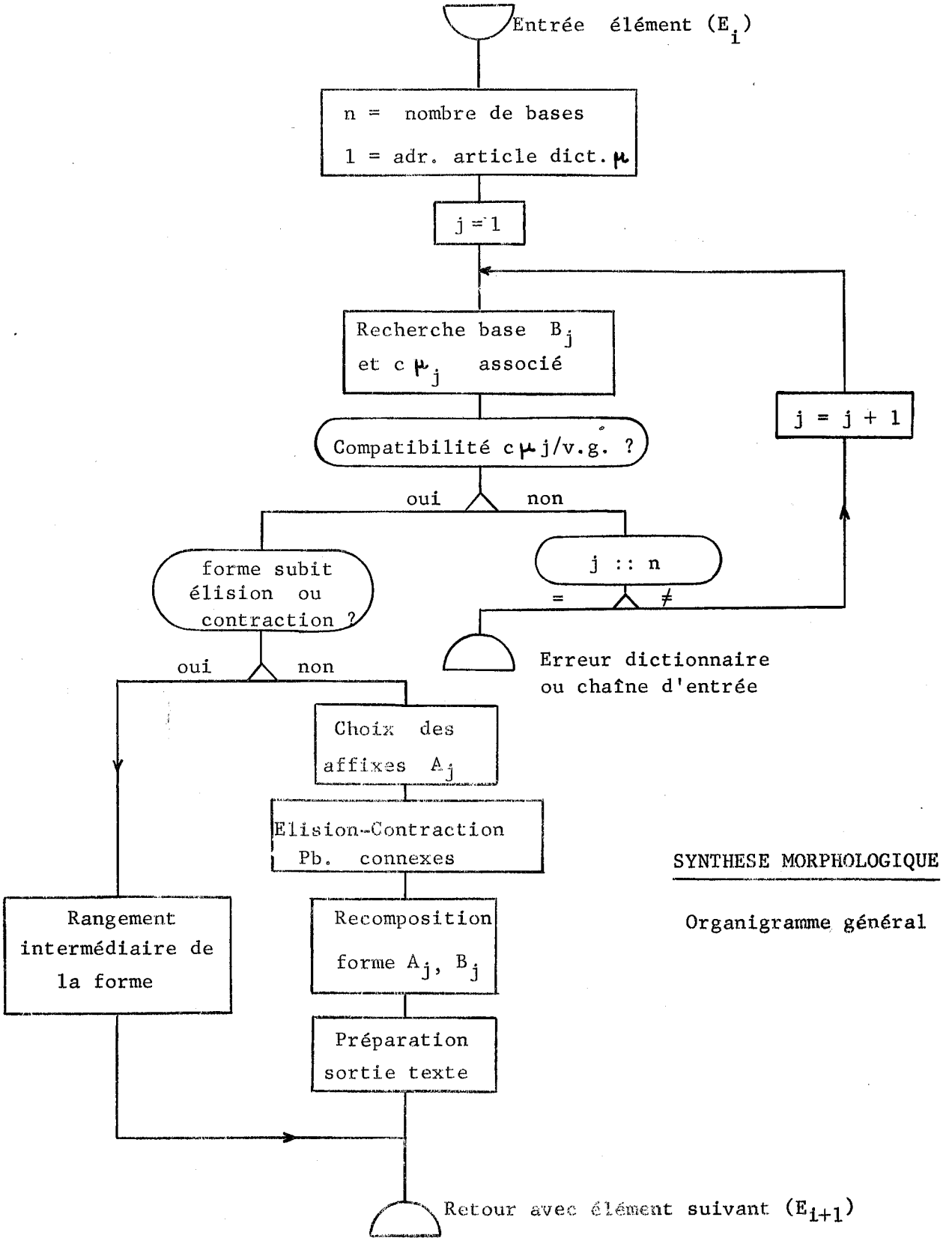
Nous donnons ci-après l'organigramme d'ensemble du programme de synthèse morphologique.

IV - UTILISATION D'UNE TECHNOLOGIE A DISQUES

Comme on peut le voir sur le schéma d'organisation générale de la synthèse, nous avons utilisé les bandes pour supporter l'information ; il est nécessaire de procéder à des rebobinages et

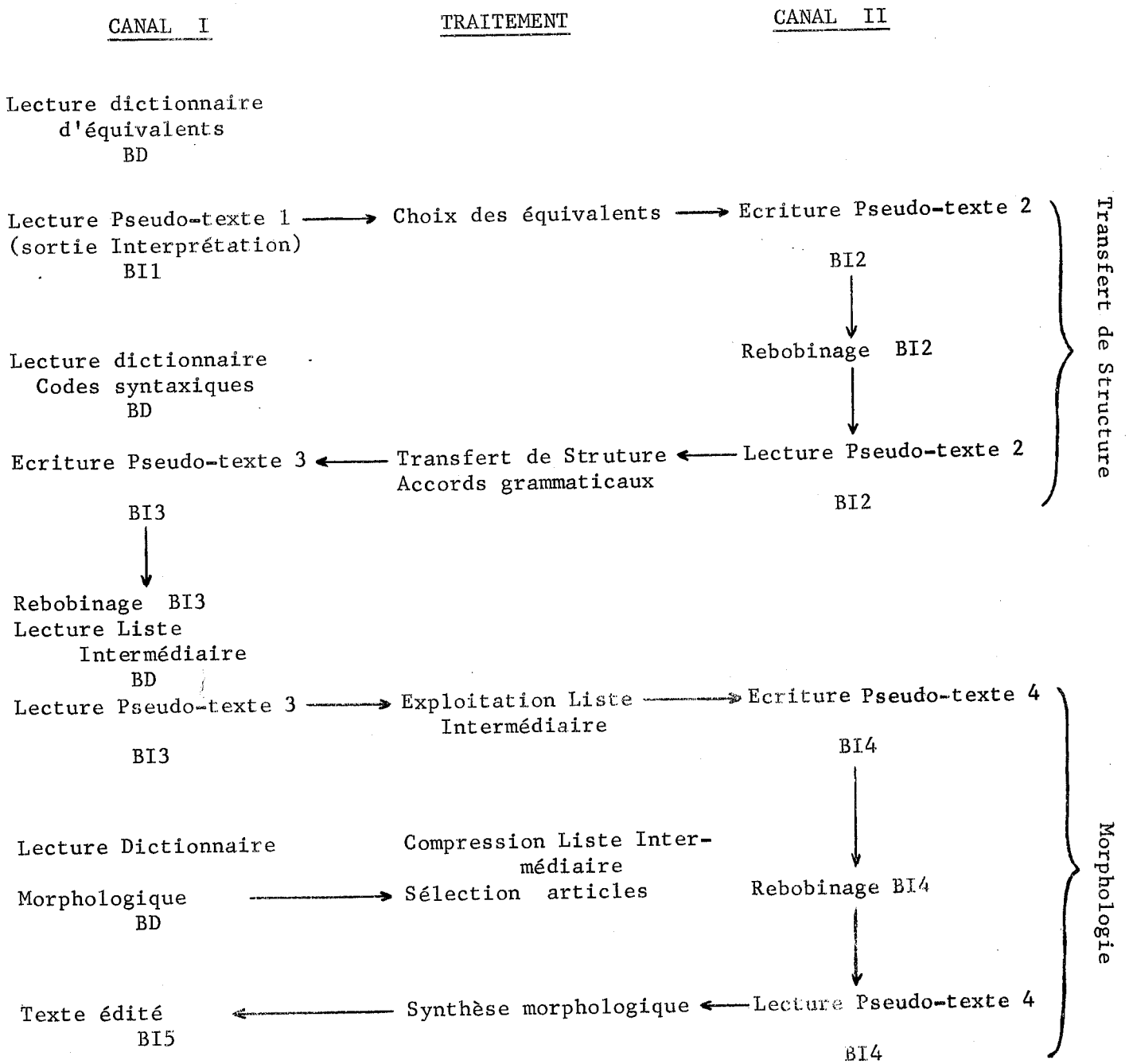
à une recherche séquentielle des informations. La technologie à disques supprime cette inconvénient ; en particulier, les consultations de dictionnaire pourront être simplifiées, la liste intermédiaire partiellement supprimée. Le temps d'accès à une position donnée est compris entre 50 et 180 millisecondes (disques 1301).

Cependant, l'ordre général des phases successives sera conservé, les disques modifiant partiellement la stratégie.



SYNTHESE MORPHOLOGIQUE

Organigramme général



ORGANISATION GENERALE ET DEROULEMENT DE LA SYNTHES

(BD ⟶ bande dictionnaire)
(BI_j ⟶ bandes intermédiaires)

Remarque : Une procédure de reprise est prévue quand le dictionnaire morphologique est trop volumineux.

CHAPITRE VI

R E A L I S A T I O N S P R A T I Q U E S

I - PROCEDURE AUTOMATIQUE DE CONJUGAISON DES VERBES FRANCAIS

Dans le cadre de l'étude morphologique de la langue française, nous avons été conduit à découper les formes en préfixe, base et affixe. Dans ce cas particulier, nous avons considéré seulement les temps simples, l'infinitif, le participe présent et le participe passé des verbes ; ceci produit un découpage en base et désinence seulement. Pour vérifier ce découpage, le programme considéré génère automatiquement toutes ces formes pour tous les verbes contenus dans le dictionnaire d'essai. Ce dernier se présente sous un volume très réduit et contient tous les types de découpages qui peuvent être rencontrés en français [20] . Le dictionnaire peut être augmenté en respectant la codification décrite ci-après.

A - Renseignements linguistiques

1 - Bases

Une base suffit pour les verbes français dits "réguliers" qui constituent 80 % de l'ensemble des verbes. Pour les autres, dits "irréguliers", nous aurons plusieurs bases (six au maximum).

2 - Désinences

Les désinences des temps simples sont regroupées par systèmes de désinences ; certains systèmes seront incomplets c'est à dire qu'ils ne contiendront les terminaisons que d'un ou plusieurs temps (ou même d'une fraction de temps). Il existe six systèmes complets et neuf incomplets, soit 330 désinences pour les temps simples.

Ainsi à chaque base du dictionnaire sera associé un seul numéro de système de désinences.

Cinq terminaisons ont été distinguées pour l'infinitif, trois pour le participe présent et sept pour le participe passé (y compris la terminaison nulle car la base peut correspondre à la forme du participe passé).

B - Composition du code morphologique

1 - Formalisation

A chaque base est associé un code morphologique occupant trente six positions binaires (une mémoire). On y trouvera les informations suivantes :

a) Possibilité d'élision (désignée par E)
b) Numéros de terminaison pour l'infinitif, le participe présent et le participe passé (δ^{I}, δ^{PR} et δ^{PA}).
c) Numéro de système de désinences pour les temps simples ($S\delta$).
d) "Capacité" de la base, c'est à dire l'indication que la base permettra d' obtenir la conjugaison du verbe à telle personne de tel temps.

L'étude linguistique a montré qu'il n'était pas nécessaire de distinguer toutes les personnes de tous les temps simples et qu'il suffisait de différencier treize classes regroupant un certain nombre de formes verbales :

1 - Indicatif Présent	Singulier
2 - Indicatif Présent	1ère et 2e personne du pluriel
3 - Indicatif Présent	3e personne du pluriel
4 - Indicatif Imparfait	
5 - Passé Simple	
6 - Futur Simple	
7 - Impératif Présent	Singulier
8 - Impératif Présent	Pluriel
9 - Conditionnel Présent	
10- Subjonctif Présent	Singulier
11- Subjonctif Présent	1ère et 2e personne du pluriel
12- Subjonctif Présent	3e personne du pluriel
13- Subjonctif Imparfait	

Par exemple, pour le verbe ALLER qui comporte plusieurs bases, la base ALL aura la capacité suivante :

A L L 0 0 0 1 1 0 0 1 0 0 1 0 1

c'est à dire que les formes appartenant aux classes 4, 5, 8, 11, et 13 seront construites à partir de cette base et de Sδ.

Pour un verbe donné, le résultat de la disjonction des "capacités" attribuées à chaque base doit comporter des "1" pour les treize classes ; dans le cas contraire soit il y a erreur d'indexage, soit le verbe est défectif.

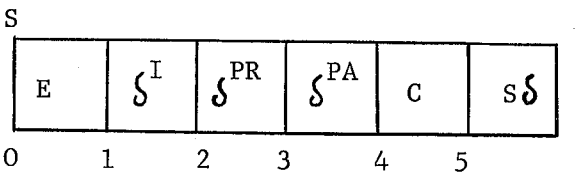
On dit que deux bases (appartenant à des verbes différents) ont le même numéro de catégorie si, et seulement si, le résultat de la disjonction de leur capacité respective est nul. La catégorie sera désignée par C .

Le code morphologique (c μ) est donc défini par

$$<c\ \mu> ::= <E><\delta^{I}><\delta^{PP}><\delta^{PA}><C><S\delta>$$

2 - Code morphologique machine

La disposition des informations dans la mémoire est indiquée ci-dessous :



- E occupe la position signe de la mémoire
- δ^{I} , δ^{PR} et δ^{PA} sont placés dans les caractères 1, 2 et 3 respectivement. Si la base n'est pas utilisée pour obtenir l'une de ces formes, alors le caractère correspondant est nul.

- Théoriquement, le nombre de catégories est 2^{13} (nombre de combinaisons différentes de 1 et de 0 sur les 13 classes) ; en fait, on dénombre, expérimentalement, quarante configurations distinctes. Le code "capacité" sera donc remplacé par un numéro de catégorie qui se trouvera dans le caractère 4 . Une table annexe permettra de décoder cette information ; ainsi, pour la base ALL, considérée ci-dessus, C sera égal à 11. La numérotation des catégories débute à zéro ; une base non utilisée pour la conjugaison des temps simples aura C = 0 , mais δ^{I} ou δ^{PR} ou δ^{PA} (ou une combinaison quelconque des trois) aura une valeur différente de zéro.

- Le caractère 5 indique le numéro de système de désinences associé.

Si C est nul, alors Sδ sera nul.

Par exemple, pour la base ALL le code morphologique machine a la configuration suivante -, 1, 1, 1, 11, 1

C - Dictionnaire

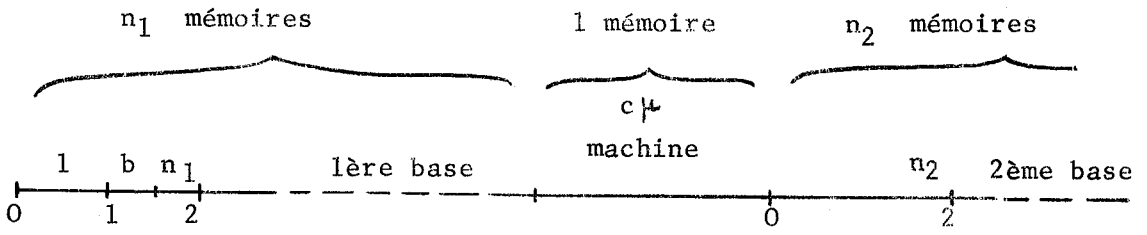
Chaque verbe (une unité lexicale) nécessite une ou plusieurs bases. Le dictionnaire comporte un article par unité lexicale.

<article de dict.> ::= <l><b> <n_1> <B_1> <cμ_{B1}> <n_2> <B_2> ...

<l> ::= longueur de l'article (nombre entier de mémoires)
<b> ::= nombre de bases appartenant à cette unité lexicale
<n_i> ::= nombre de mémoires occupées par la base B_i

On doit avoir $l = n_1 + n_2 + \ldots\ldots + n_b + b$

D'où la représentation machine



Le dictionnaire d'essai comporte tous les types de verbes selon la classification du Bescherelle, vu sa faible importance il est entièrement en mémoire rapide. Les verbes sont classés par ordre alphabétique, à l'exception des auxiliaires 'avoir' et 'être' qui sont placés en tête.

D - Déroulement du programme

Les verbes sont conjugués successivement à tous les temps précisés ci-dessus. Une mémoire du programme simule les données, c'est à dire donne l'adresse de l'article de dictionnaire et les valeurs des variables grammaticales (temps, personne, nombre, mode). Ces dernières permettent de choisir la base convenable. Si aucune base n'est satifaisante, le verbe est défectif, et ce message remplace la (ou les) forme qui n'existe pas.

Ce programme a été écrit et mis au point sur machine IBM 7090. L'ensemble des conjugaisons est sorti sur bande magnétique, celle-ci étant imprimée avec la présentation adéquate à l'aide d'un programme exécutable sur IBM 1401 (voir Annexe 1).

La conjugaison complète des 65 verbes du dictionnaire dure environ 12 secondes.

Remarque : - Programme de démonstration

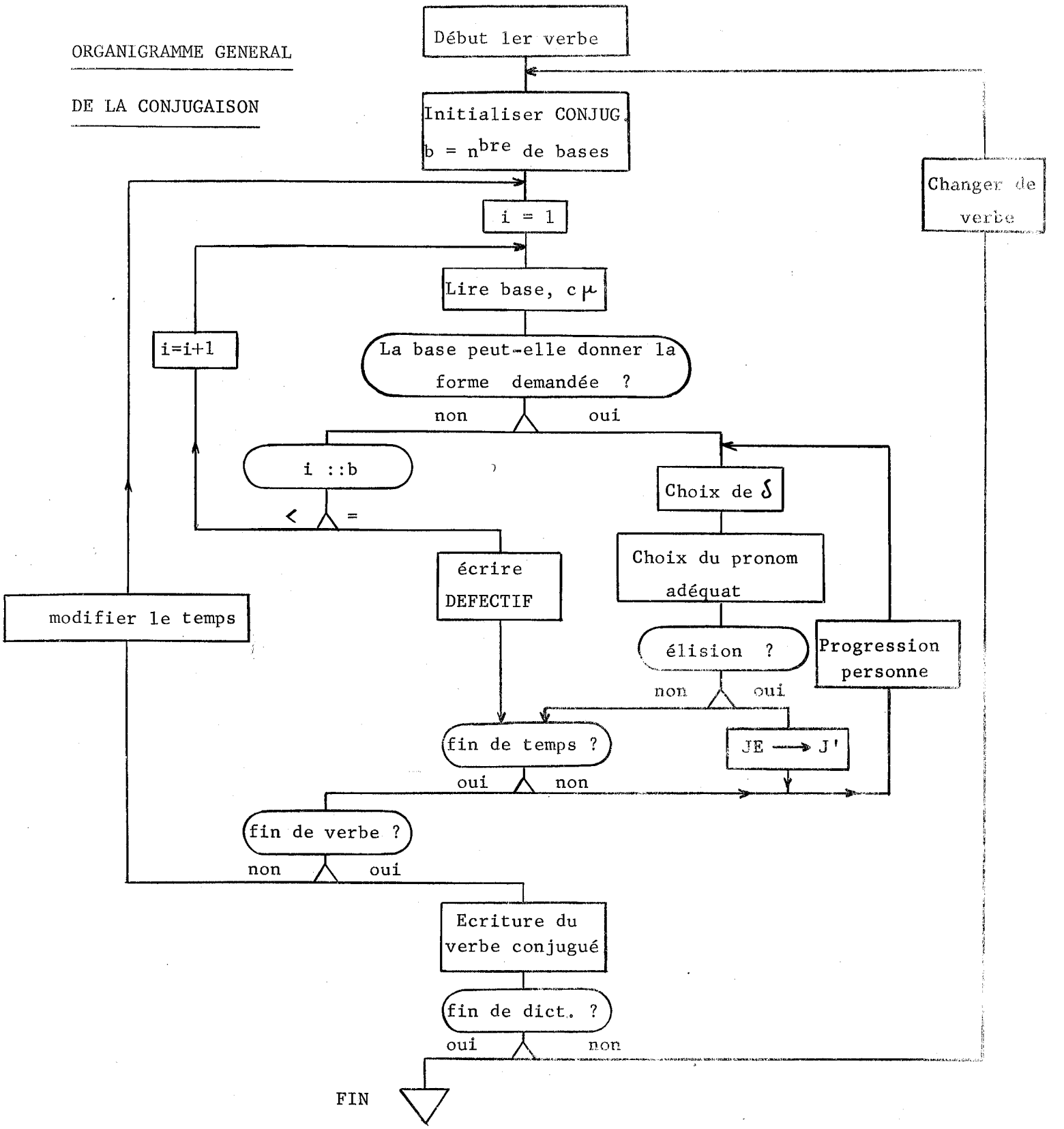
Le programme général a été modifié de façon à fournir, à la demande, la conjugaison entière, les formes d'un temps ou d'une personne d'un temps pour un verbe donné.

La question doit être perforée sur une carte selon le format indiqué en Annexe 1. Sur IBM 7090 la réponse est imprimée immédiatement sur l'imprimante connectée, la carte portant la question étant lue au lecteur de cartes.

Ce programme a été rendu compatible avec l'ordinateur IBM 7044. Le résultat est alors imprimé sur la machine à écrire (Annexe 1).

ORGANIGRAMME GENERAL

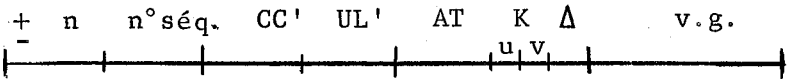
DE LA CONJUGAISON



II - ESSAI DE SYNTHESE MORPHOLOGIQUE *

Cet essai porte uniquement sur la phase terminale du processus de synthèse, c'est à dire que le dictionnaire morphologique et le pseudo-texte sont supposés être en mémoire rapide.

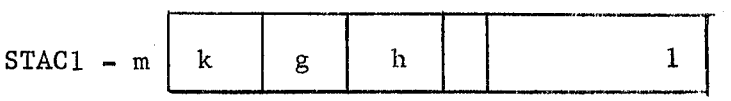
Le dictionnaire respecte les conventions énoncées aux chapitres IV et V. Arbitrairement, nous avons fixé un format pour le pseudo-texte d'entrée : quatre mémoires par élément.



Si le signe est - , forme non alphabétique ; n est le nombre de bases. La disposition des variables grammaticales pour chaque classe est donnée en annexe.

A - Recherche d'une base et de son code morphologique

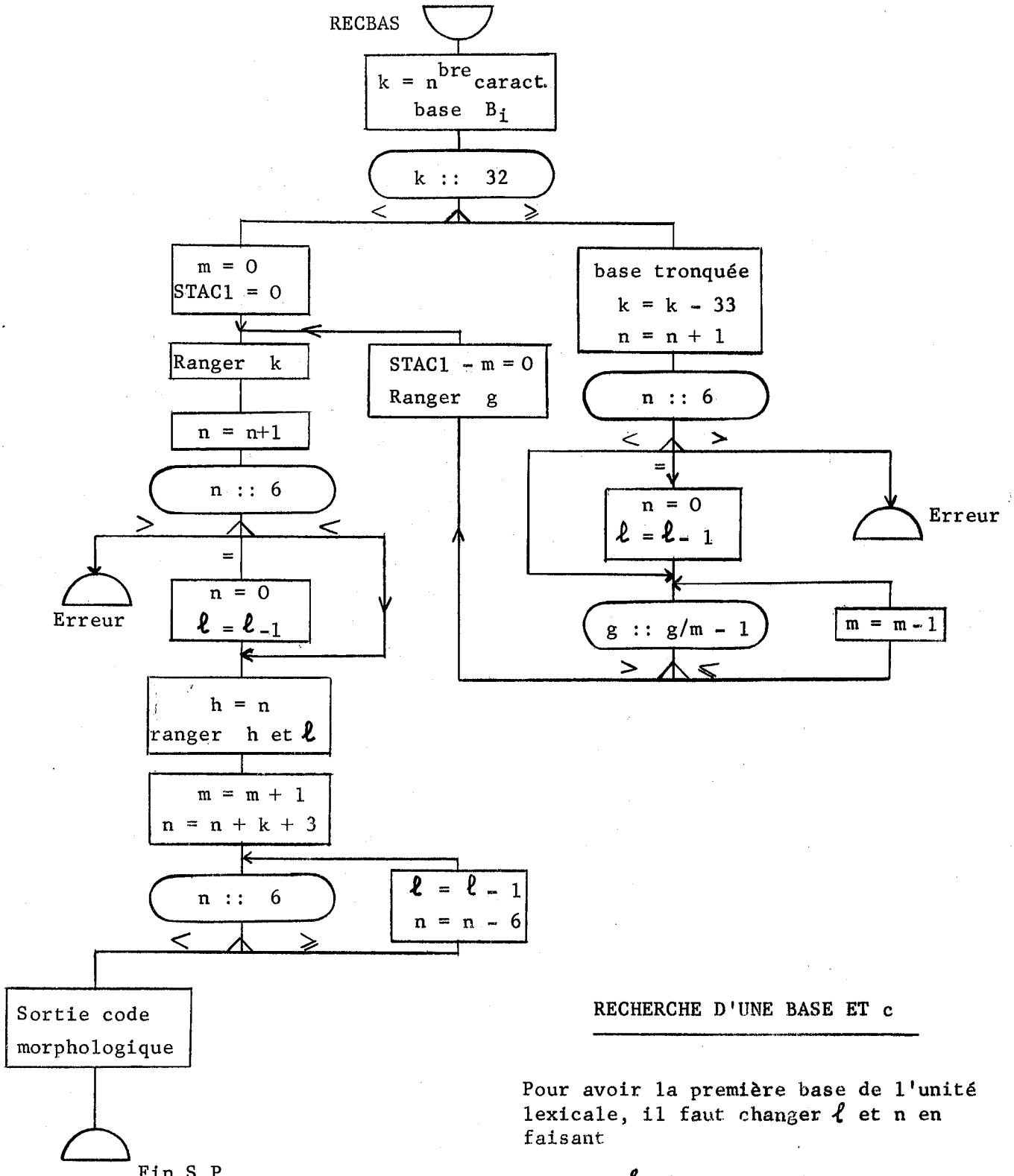
Si une base est tronquée, selon le procédé expliqué ci-dessus, il faut conserver les renseignements sur les bases précédentes pour effectuer la recomposition de la forme. Le sous-programme, qui délivre le code morphologique associé à la base, est indépendant de la classe morphologique. Il conserve les informations spécifiques à la base elle-même dans un "réservoir" (appelé STAC1), dans celui-ci une mémoire est associée à chaque base, avec l'indice m :



Le premier caractère de la base est situé au caractère h de la mémoire d'adresse l ; cette base a une longueur de k caractè-

---

* Le traitement de la classe VERBE a été écrit et mis au point par Mme BOUSSARD.



RECHERCHE D'UNE BASE ET c

Pour avoir la première base de l'unité lexicale, il faut changer ℓ et n en faisant

$$\ell = AT \text{ et } n = 0$$

Ensuite ces valeurs sont conservées dans le S.P.

res, et elle a g caractères communs avec la base décrite en STAC1 - (m-1). Par exemple, on commande la recherche de B_i ; si celle-ci est une base tronquée, on compare h_i et h_{m-1} ; si $h_i \leqslant h_{m-1}$, il n'y a pas lieu de conserver la description de la base correspondant à (m-1) et ainsi de suite ... (voir organigramme).

B - Recomposition des formes

On a vu que les formes ont principalement, les deux types suivants (voir chapitre IV)

$$P\ B\ \delta_1\ S\ \delta_2\ \delta_3$$

$$B'\ \delta'_1\ \delta'_2\ \lambda'\ B\ \delta_1\ \delta_2$$

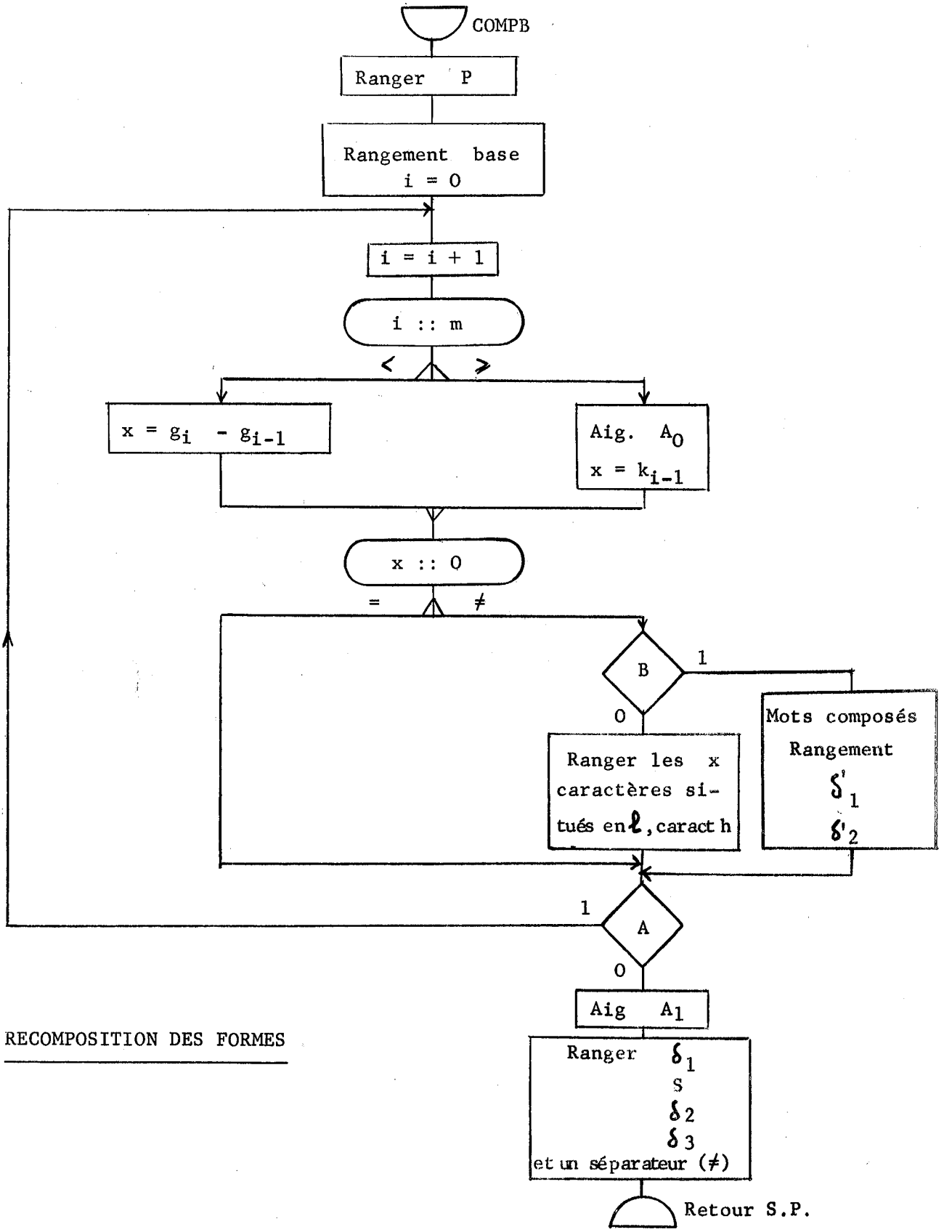
Le sous-programme de recomposition prend ces composants et les met dans l'ordre donné. Pour la base, qui peut être comprimée, les renseignements notés lors de la recherche de la base sont utilisés. Par ailleurs, il faut faire la distinction entre les formes de type 1 et de type 2 (mots composés). Un séparateur (blanc) est introduit à la fin de chaque forme.

C - Elision - Contraction

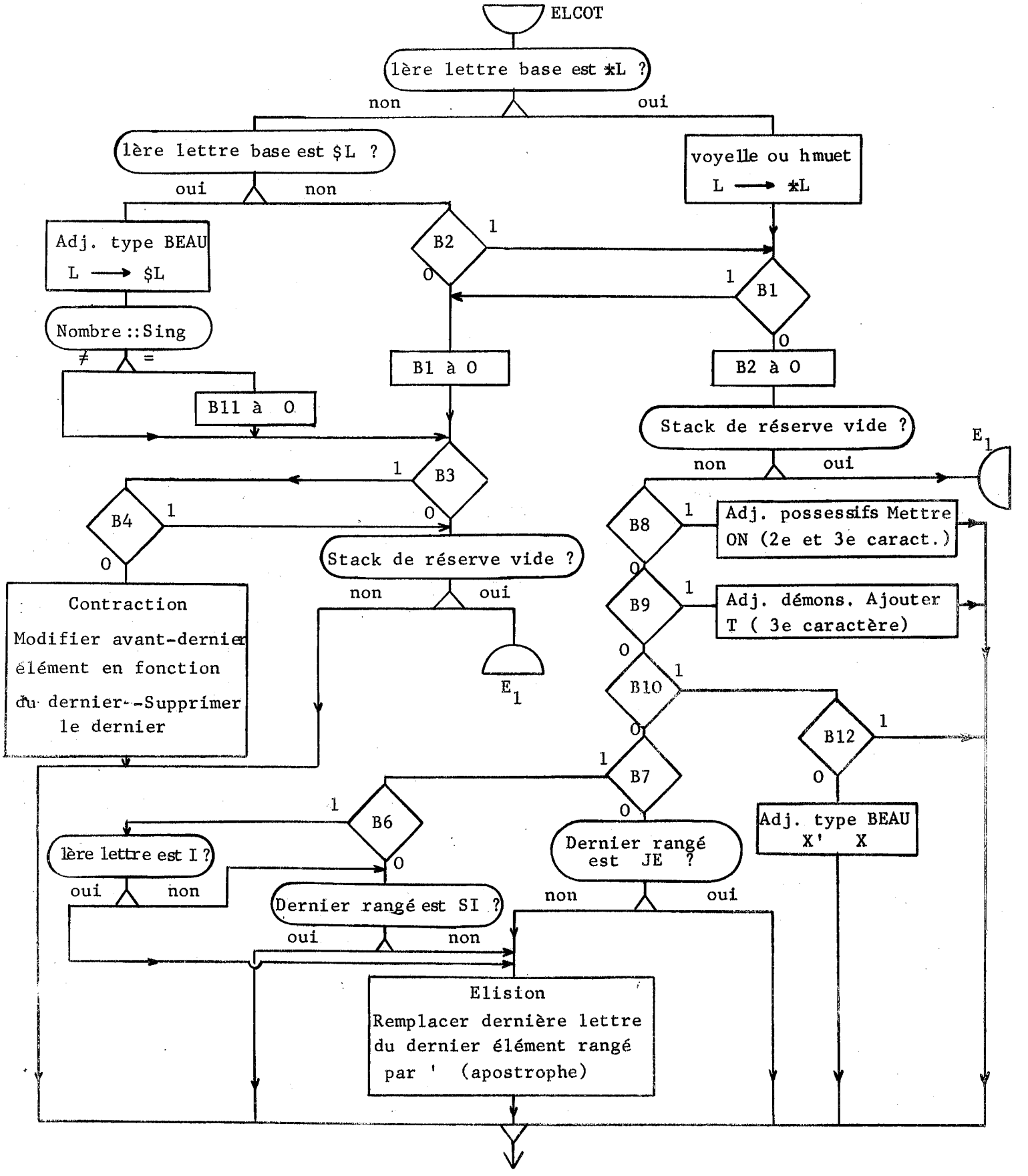
Nous donnerons seulement l'organigramme de traitement. Remarquons que chaque forme est susceptible de provoquer l'élision, car un préfixe commençant par une voyelle peut être ajouté dans de nombreux cas.

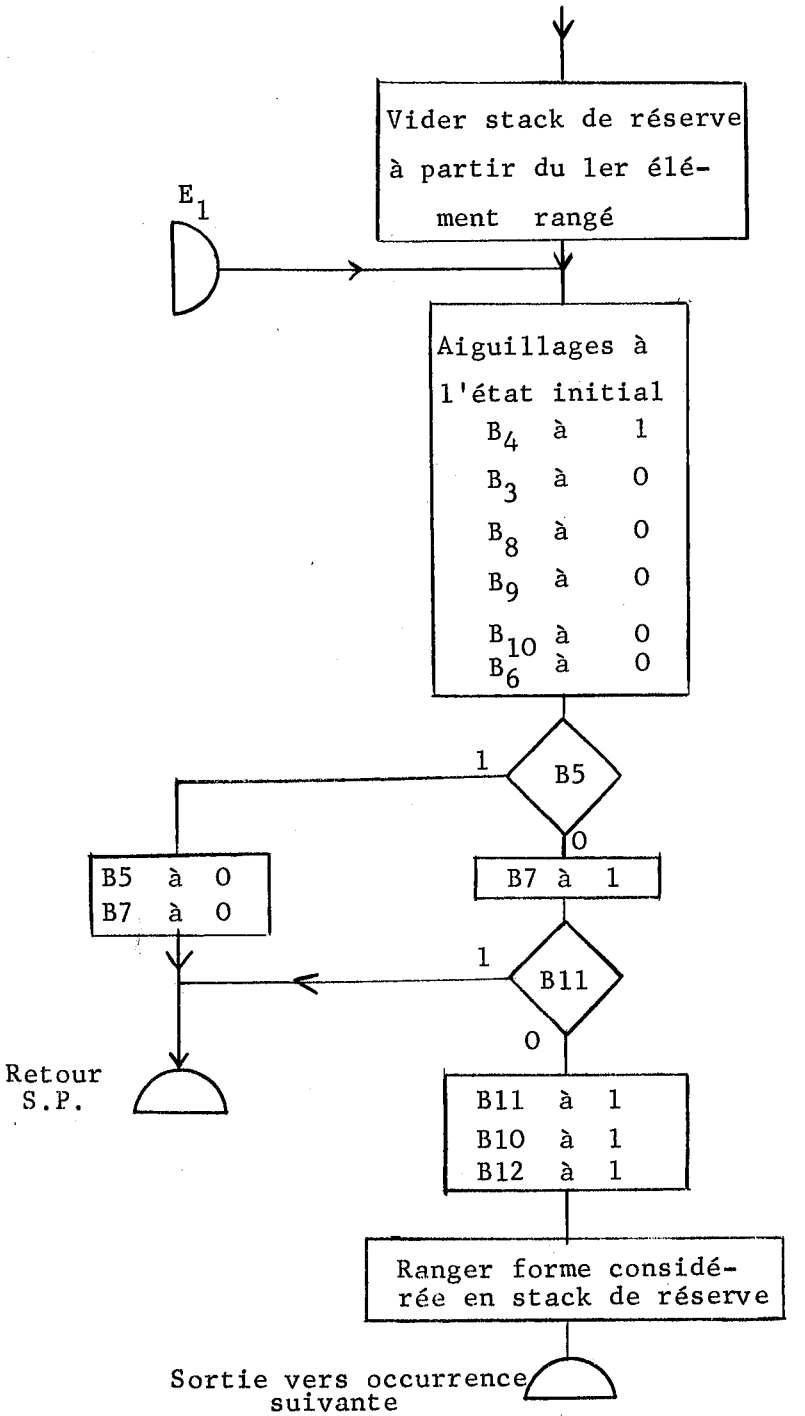
D - Résultats obtenus

Ce programme fut d'abord mis au point sur ordinateur



RECOMPOSITION DES FORMES


ELCOT
1ère lettre base est *L ?
non
oui
1ère lettre base est $L ?
voyelle ou hmuet
L → *L
oui
non
Adj. type BEAU
L → $L
B2
1
0
1
B1
0
Nombre ::Sing
≠
=
B1 à 0
B2 à 0
B11 à 0
Stack de réserve vide ?
E1
non
oui
1
B3
0
B4
1
0
B8
1
Adj. possessifs Mettre ON (2e et 3e caract.)
0
B9
1
Adj. démons. Ajouter T (3e caractère)
Stack de réserve vide ?
0
B10
1
non
oui
E1
Contraction
Modifier avant-dernier élément en fonction du dernier--Supprimer le dernier
0
B12
1
1
B7
0
1
B6
0
Adj. type BEAU
X' X
1ère lettre est I ?
Dernier rangé est JE ?
oui
non
non
oui
Dernier rangé est SI ?
oui
non
Elision
Remplacer dernière lettre du dernier élément rangé par ' (apostrophe)



ELISION

CONTRACTION

PROBLEMES CONNEXES

Les aiguillages sont positionnés par les composants de la forme et par la ou les formes précédentes. Par exemple :

- B_2 est mis à 1 par les préfixes commençant par une voyelle.
- B_4 est mis à 0 par les prépositions A ou DE.

IBM 7090 pour la synthèse du substantif. Il a été entièrement refondu sur IBM 7044 pour traiter toutes les classes morphologiques - à l'exception de l'adjectif cardinal - Un exemple est donné en Annexe.

III - PROGRAMMES DE SERVITUDE

Ces programmes peuvent se diviser en deux catégories :

- Programmes d'aide aux linguistes
- Programmes de servitude d'exploitation (dictionnaire, etc ...).

Les documents fournis par les linguistes consistent en feuilles d'indexage et feuilles d'équivalents ; les programmes exploitent ces informations et leur donnent un format fixe assimilable par le calculateur.

A - Programmes d'aide aux linguistes

1 - Dictionnaire des formes conomiques

Cette liste est obtenue à partir de la feuille d'indexage fournie par le linguiste :

Base	n° UL, CC n° base nbre bases	Renseignements morphologiques : $c\,\mu_1$	Dérivations
1	2	3	4

Nous nous limiterons aux quatre classes morphologiques dont nous définissons ci-dessous les formes canoniques :

F. C. Substantif : Forme au Masculin - Singulier
F. C. Adjectif : Forme au Masculin - Singulier
F. C. Verbe : Forme infinitive
F. C. Adverbe : Forme adverbe

Pour chaque forme canonique nous aurons un article de dictionnaire :

$$\langle \text{article dict. F.C.} \rangle ::= \langle \text{F.C.} \rangle \underbrace{\langle \text{UL} \rangle \langle \text{CC} \rangle \langle \Delta \rangle}_{\text{origine}} \underbrace{\langle K_{u,v} \rangle}_{\text{résultat}} \underbrace{\langle \text{Commentaires} \rangle}_{\text{quelconques}}$$

Le programme délivre toutes les formes canoniques qui sont susceptibles d'être obtenues à l'aide des renseignements notés sur la feuille d'indexage ; les articles correspondants sont rassemblés sur bande magnétique (BFC_0).

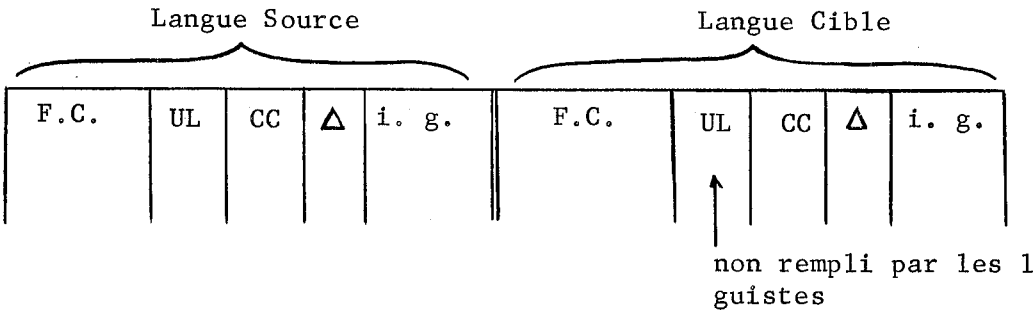
Nous trierons ces articles selon les critères suivants :

- "u" du $K_{u,v}$
- Ordre alphabétique des F.C.

avec sortie " en clair" des articles qui ne se distinguent pas sur ces deux critères, ce qui permet aux linguistes de vérifier l'indexage. Le dictionnaire des F.C. est imprimé lorsque les corrections ont été faites ; la mise à jour est effectuée par tri et interclassement (BFC_1) .

2 - Achèvement des feuilles d'équivalents

Les feuilles d'équivalents se présentent sous le format suivant :



Il y a mise sur bande magnétique des feuilles d'équivalents (BEQ_0).

On fait un tri sur les F.C. cible et on imprime celles qui sont identiques (à vérifier par les linguistes). Lorsque les corrections sont faites, cette bande remplace les feuilles d'équivalents (BEQ_1).

A partir de BEQ_0 , par tri sur :

- Code classe (cible)
- Forme canonique (cible)
- Δ (cible)

on obtient une bande BEQ_2 , sur laquelle on va porter, automatiquement, les numéros d'unité lexicale (cible) à l'aide de la bande BFC_2 obtenue (à partir de BFC_0) par tri sur

- Code classe
- Formes canoniques
- Δ

Lorsqu'il y a ambiguïté, c'est à dire lorsque plusieurs articles de BFC_2 ont les mêmes caractéristiques, le choix est laissé au linguiste qui décidera de l'équivalent (BEQ).

B - Programmes de servitude d'exploitation

1 - Dictionnaire morphologique

Nous avons vu sa description dans les chapitres précédents ; il est obtenu à partir des feuilles d'indexages : on supposera que celles-ci sont triées par code classe et n° d'unité lexicale. Un certain nombre de vérifications seront faites lors de la fabrication de la bande dictionnaire (BDM) : cohérence des codes, etc ...

2 - Liste intermédiaire

La liste intermédiaire est construite selon les indications données précédemment. Une mémoire est affectée à chaque unité lexicale. Elle peut être fabriquée en même temps que le dictionnaire morphologique (BLI), car on a besoin des renseignements contenus sur les feuilles d'indexage.

3 - Dictionnaire syntaxique

Une mémoire est associée à chaque unité lexicale et contient le code syntaxique, cette mémoire a la même adresse que celle associée à la liste intermédiaire. Il peut être fabriqué à part , les codes syntaxiques étant triés selon les mêmes critères que les feuilles d'indexage (BDS).

4 - Dictionnaire d'équivalents

Un article de la bande BEQ est de la forme :

$$<\text{article BEQ}> ::= <FC_S> <UL_S , CC_S><\Delta_S> <ig_S> <FC_C>$$

$$<UL_C , CC_C> <\Delta_C> <ig_C>$$

(l'indice S désigne la langue source)

Ces articles sont triés sur CC_C , UL_C (BDE_O) et le programme de fabrication de la liste intermédiaire placera dans chaque article l'adresse A' adéquate (BDE1).

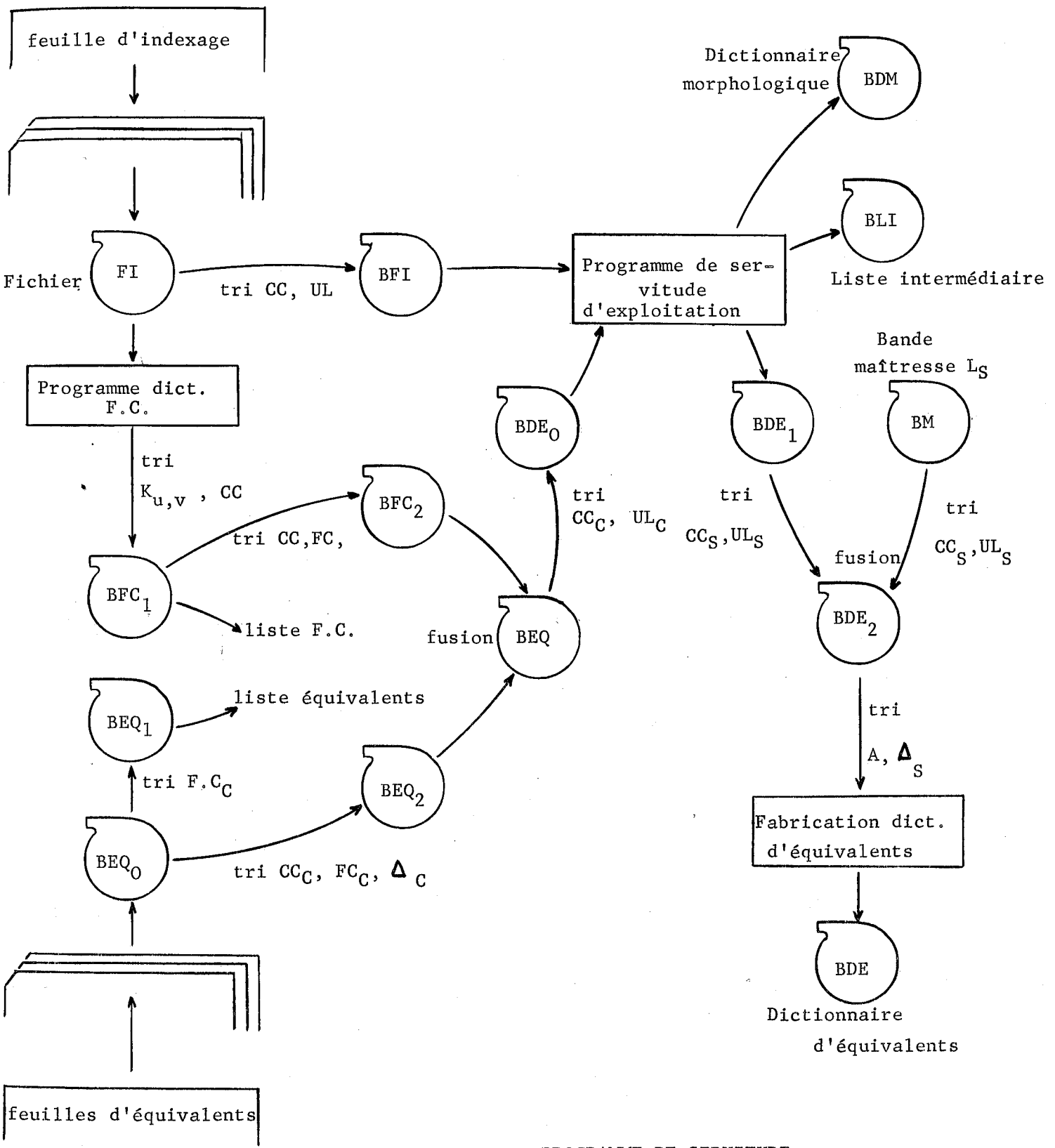
A partir de la bande maîtresse, obtenue en langue source, qui est de la forme :

$$<A><\text{Base}> <CC_S , UL_S ; n° \text{ base}> <c\mu>$$

on insère sur BDE_1 l'adresse A obtenue en langue source. Les deux bandes ayant été triées sur CC_S , UL_S , la bande résultat BDE_2 sera triée sur

- les adresses A
- la dérivation Δ_S

Nous obtiendrons ainsi le dictionnaire d'équivalents (BDE).



PROGRAMME DE SERVITUDE

CONCLUSION

Cette étude a permis de dégager un certain nombre de problèmes spécifiques à la synthèse d'une langue. A quelques problèmes, il a pu être apporté une solution complète (consultation des dictionnaires). Par contre certains points n'ont pu être développés car il manque actuellement les études linguistiques qui conditionneront l'application de la théorie ; c'est le cas en particulier des règles de transfert de structure.

Le canevas ainsi tracé ne dépend pas des calculateurs utilisés ; nous avons supposé cependant une grande capacité de mémoires à accès rapide, et des mémoires à accès séquentiel (bandes). L'évolution des techniques permet de prévoir une augmentation de la capacité jointe à une diminution du temps d'accès, ce qui ne peut que favoriser la résolution des différents problèmes posés par la traduction automatique.

BIBLIOGRAPHIE

1 - B. VAUQUOIS - Langages artificiels, systèmes formels et Traduction Automatique
Nato Advanced Study - Venise juillet 62

2 - J. BLOIS, E. MORLET - Morphologie du français pour la traduction automatique
Rapport CETIS n° 44 - avril 1962

3 - S. LAMB, W. JACOBSON - A high-speed capacity Dictionnary System
Mechanical Translation - Vol 6. Nov. 61
p. 76 - 107

4 - V. YNGVE - A framework for Syntactic Translation
Mechanical Translation - Vol 4 - n°3 1959

5 - V. YNGVE - A model and a hypothesis for Language Structure
Proc. Amer. Phil. soc. - 1960 104, n°5 440

6 - S. CECCATO, MARETTI, ZONTA - Linguistic Analysis and Programming for M. T.
Report RADC - TR - 60 - 18 - Université de Milan

7 - G. SALTON - Manipulation of trees for Information Retrieval
Communications of the ACM - février 1962
Vol. 5. n° 2

8 - K. IVERSON - A Programming Language
Wiley - New York (1962)

9 - G. VEILLON - Consultation d'un dictionnaire et analyse morphologique en Traduction Automatique
Thèse de 3ème cycle. Université de Grenoble - Juin 1962

10 - G. VEILLON - Présentation du programme de dictionnaire allemand
CETA - document G500 A

11 - B. VAUQUOIS, J. VEYRUNES - Présentation de l'analyse morphologique du Russe
CETA - Document G 100 - C

12 - A. AUROUX - Contribution à la reconnaissance des structures syntaxiques en Traduction Mécanique
Thèse de 3ème cycle. Université de Grenoble
Juin 1962

13 - D. AUGEREAU - Utilisation des informations sémantiques en Traduction Automatique
Thèse de 3ème cycle, Université de Grenoble - Octobre 1962

14 - V. MARMA - The Machine Representation of abstract trees for an information storage and retrieval system.
Report ISR - 1 - Harvard Computation Laboratory 1961.

15 - M. GROSS - On the equivalence of models of language used in the fields of M. T. and I.R.
NATO Advanced Study Institute - Venise - juillet 1962

16 - M. GROSS - Linguistique Mathématique et langages de programmation
Revue AFCALTI - 1963 - n°4

17 - N. CHOMSKY - Formal properties of Grammars
Handbook of Mathematical Psychology - 1963

18 - G. VEILLON, J. VEYRUNES - Présentation des programmes d'Entrées - Sorties utilisés au CETA sur IBM 7090
CETA - Document C - TC - 9

19 - F. RONSSE - Locutions françaises issues de mots russes uniques.
Université Librede Bruxelles

20 - L'art de conjuguer, le nouveau Bescherelle.

21 - W. D. FOUST, J. R. WALKLING - A preliminary structural transfer system.
First International Conference on Machine Translation, Sept. 61 - Teddington, England.

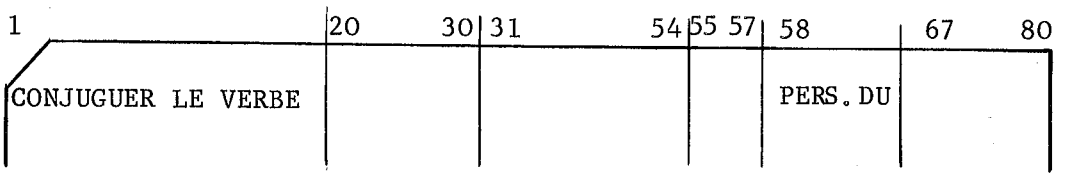
22 - E. SUSSENGUTH, JR. - Automatic Structure - Matching Procedures
Report ISR - 5 - Harvard Computation Laboratory 1964.

23 - A. BOUSSARD, M. BERTHAUD - Présentation de la synthèse morphologique du Français - CETA Document G1500 - A

24 - D. G. HAYS - Grouping and dependency Theories 1960 - Rand Corporation - P 1960 Santa Monica - California

A N N E X E 1

Format de la carte perforée correspondant à une question posée au programme de démonstration (conjugaison des verbes français).



Colonnes 20 - 30 : Infinitif du verbe, cadré en 20

Colonnes 31 - 54 : Temps et mode (selon la liste donnée page suivante) cadré en 31.

Colonne 55 : Personne (1, 2 ou 3)

A partir de la colonne 67 : nombre (SING. ou PLUR.)

Si on demande :

- La conjugaison complète d'un verbe : carte vierge à partir de 31.
- Un temps : carte vierge à partir de 55

Un message est imprimé si on a demandé un verbe qui n'est pas dans le dictionnaire ou si la question est incohérente.

CONJUGAISON AUTOMATIQUE DES VERBES

I - LISTE DES VERBES ACTUELLEMENT DANS LE DICTIONNAIRE :

AVOIR
ETRE

ABSOUDRE
ACQUERIR
AIMER
ALLER
ASSAILLIR
ASSEOIR

BATTRE
BOIRE
BOUILLIR
BROYER

CLORE
CONCLURE
CONFIRE
CONNAITRE
COUDRE
COURIR
CRAINDRE
CROIRE
CROITRE
CUEILLIR
CUIRE

DEVOIR
DIRE
DORMIR

ECRIRE
ENVOYER

FAIRE
FINIR
FUIR

JETER
JOINDRE

LIRE

MANGER
MAUDIRE
MEDIRE
MENTIR
METTRE
MOUDRE
MOURIR
MOUVOIR

NAITRE

PAITRE
PEINDRE
PLAIRE
POURVOIR
POUVOIR
PRENDRE

RECEVOIR
RENDRE
RIRE

SAVOIR
SERVIR
SOUFFRIR
SUIVRE
SURSEOIR

TENIR
TRAIRE

VAINCRE
VALOIR
VETIR
VIVRE
VOIR
VOULOIR

II - LISTE DES TEMPS :

Indicatif Présent	Condit. Présent	Impératif Présent
Indicatif Imparfait	Subjonc. Présent	Infinitif
Passé Simple	Subjonc. Imparfait	Participe Présent
Futur Simple		Participe Passé

III - QUESTIONS POSSIBLES :

- CONJUGAISON COMPLETE,
- UN TEMPS,
- LA PERSONNE D'UN TEMPS.

C.E.T.A. GRENOBLE - PROCEDURE AUTOMATIQUE DE CONJUGAISON

CONJUGUER LE VERBE ABSOUDRE INDICATIF IMPARFAIT 2 PERS. DU SING.

TU ABSOLVAIS

CONJUGUER LE VERBE METTRE IMPERATIF PRESENT 2 PERS. DU SING.

METS

CONJUGUER LE VERBE SURSEOIR SUBJONCTIF PRESENT

QUE JE SURSOIE
QUE TU SURSOIES
QU'IL SURSOIE
QUE NOUS SURSOYIONS
QUE VOUS SURSOYIEZ
QU'ILS SURSOIENT

CONJUGUER LE VERBE MOUDRE

INDICATIF PRESENT
JE MOUDS
TU MOUDS
IL MOUD
NOUS MOULONS
VOUS MOULEZ
ILS MOULENT

INDICATIF IMPARFAIT
JE MOULAIS
TU MOULAIS
IL MOULAIT
NOUS MOULIONS
VOUS MOULIEZ
ILS MOULAIENT

PASSE SIMPLE
JE MOULUS
TU MOULUS
IL MOULUT
NOUS MOULUMES
VOUS MOULUTES
ILS MOULURENT

FUTUR SIMPLE
JE MOUDRAI
TU MOUDRAS
IL MOUDRA
NOUS MOUDRONS
VOUS MOUDREZ
ILS MOUDRONT

CONDITIONNEL PRESENT
JE MOUDRAIS
TU MOUDRAIS
IL MOUDRAIT
NOUS MOUDRIONS
VOUS MOUDRIEZ
ILS MOUDRAIENT

SUBJONCTIF PRESENT
QUE JE MOULE
QUE TU MOULES
QU'IL MOULE
QUE NOUS MOULIONS
QUE VOUS MOULIEZ
QU'ILS MOULENT

SUBJONCTIF IMPARFAIT
QUE JE MOULUSSE
QUE TU MOULUSSES
QU'IL MOULUT
QUE NOUS MOULUSSIONS
QUE VOUS MOULUSSIEZ
QU'ILS MOULUSSENT

IMPERATIF PRESENT
MOUDS
MOULONS
MOULEZ

INFINITIF
MOUDRE

CONJUGUER LE VERBE PEINDRE INFINITIF

PEINDRE

CONJUGUER LE VERBE CONFIRE PARTICIPE PRESENT

CONFISANT

CONJUGAISON DES VERBES FRANCAIS

VERBE NR. 24

INDICATIF PRESENT

JE DCIS TU DCIS IL DCIT NCUS DEVCNS VOUS DEVEZ ILS DOIVENT

INDICATIF IMPARFAIT

JE DEVAIS TU DEVAIS IL DEVAIT NCUS DEVICNS VOUS DEVIEZ ILS DEVAIENT

PASSE SIMPLE

JE DUS TU DUS IL DUT NOUS DUMES VCUS DUTES ILS DURENT

FUTUR SIMPLE

JE DEVRAI TU DEVRAS IL DEVRA NCUS DEVRCNS VOUS DEVREZ ILS DEVRCNT

CONDITIONNEL PRESENT

JE DEVRAIS TU DEVRAIS IL DEVRAIT NCUS DEVRICNS VOUS DEVRIEZ ILS DEVRAIENT

SUBJCNCTIF PRESENT

QUE JE DCIVE QUE TU DOIVES QU'IL DOIVE QUE NCUS DEVIONS QUE VOUS DEVIEZ
QU'ILS DCIVENT

SUBJCNCTIF IMPARFAIT

QUE JE DUSSE QUE TU DUSSES QU'IL DUT QUE NCUS DUSSICNS QUE VOUS DUSSIEZ
QU'ILS DUSSENT

IMPERATIF PRESENT

DCIS DEVONS DEVEZ

INFINITIF

DEVCIR

A N N E X E 2

GESTION DES ENTREES - SORTIES SUR BANDES MAGNETIQUES

La bande magnétique est un support d'information à accès lent vu la rapidité des organes de calcul de l'ordinateur IBM 7090 . Sur cette machine il est possible d'effectuer en simultanéité

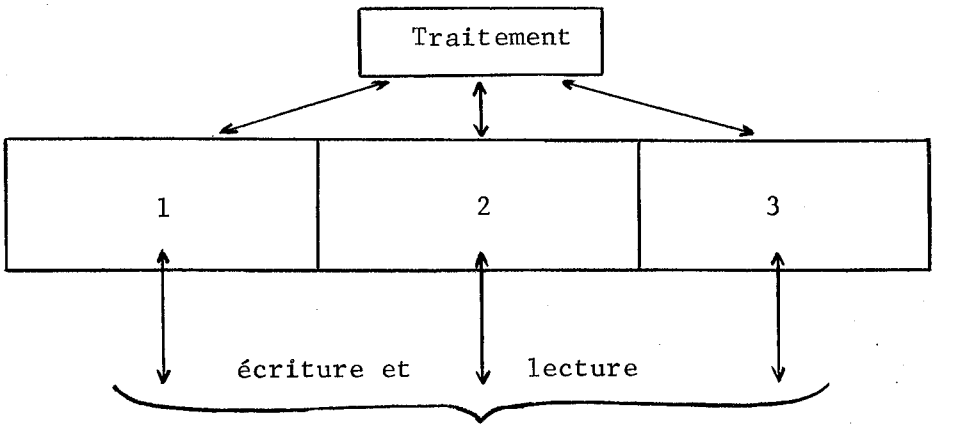
- Une lecture à partir d'une bande B (données)
- Un traitement en mémoire rapide
- Une écriture sur une bande B' (résultats)

à condition que les dérouleurs B et B' soient connectés à deux canaux différents. Le temps d'exécution sera réduit à la duré maximum de l'une des trois phases.

A - Organisation des zones d'Entrée - Sortie

Une zone de mémoires est utilisée pour la transmission des informations vers (ou à partir du) le support physique ; cette zone est divisée en "buffers" , chaque buffer étant utilisé pour transmettre un enregistrement.

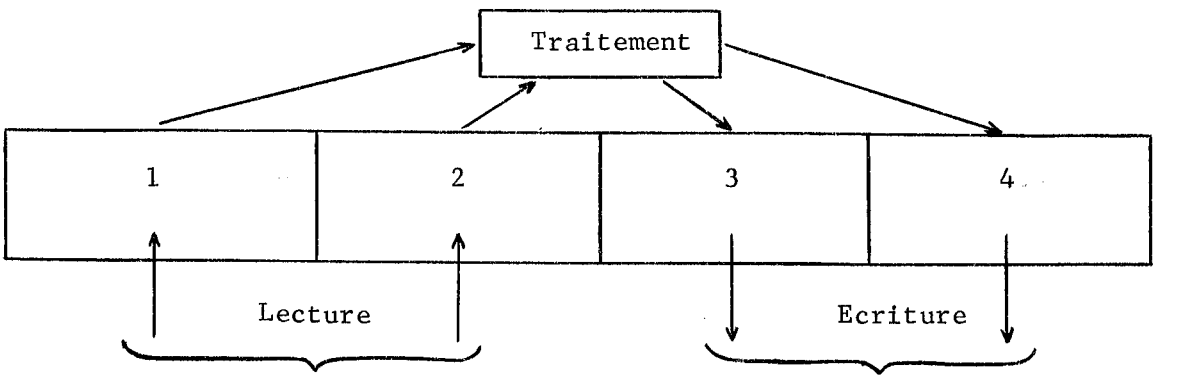
1 - La dimension des enregistrements est identique : les buffers sont banalisés



Bandes magnétiques

Le rangement des résultats traités se fait dans le buffer qui contient les données. Les deux autres buffers servent l'un à la lecture, l'autre à l'écriture.

2 - Les dimensions des enregistrements sont différentes; chaque buffer a une fonction déterminée (écriture ou lecture).



Bandes magnétiques

Remarque :

Dans les deux cas, il ne peut y avoir de simultanéité lors de la lecture du premier et l'écriture du dernier enregistrement.

B - Réalisation du programme sur ordinateur 7090

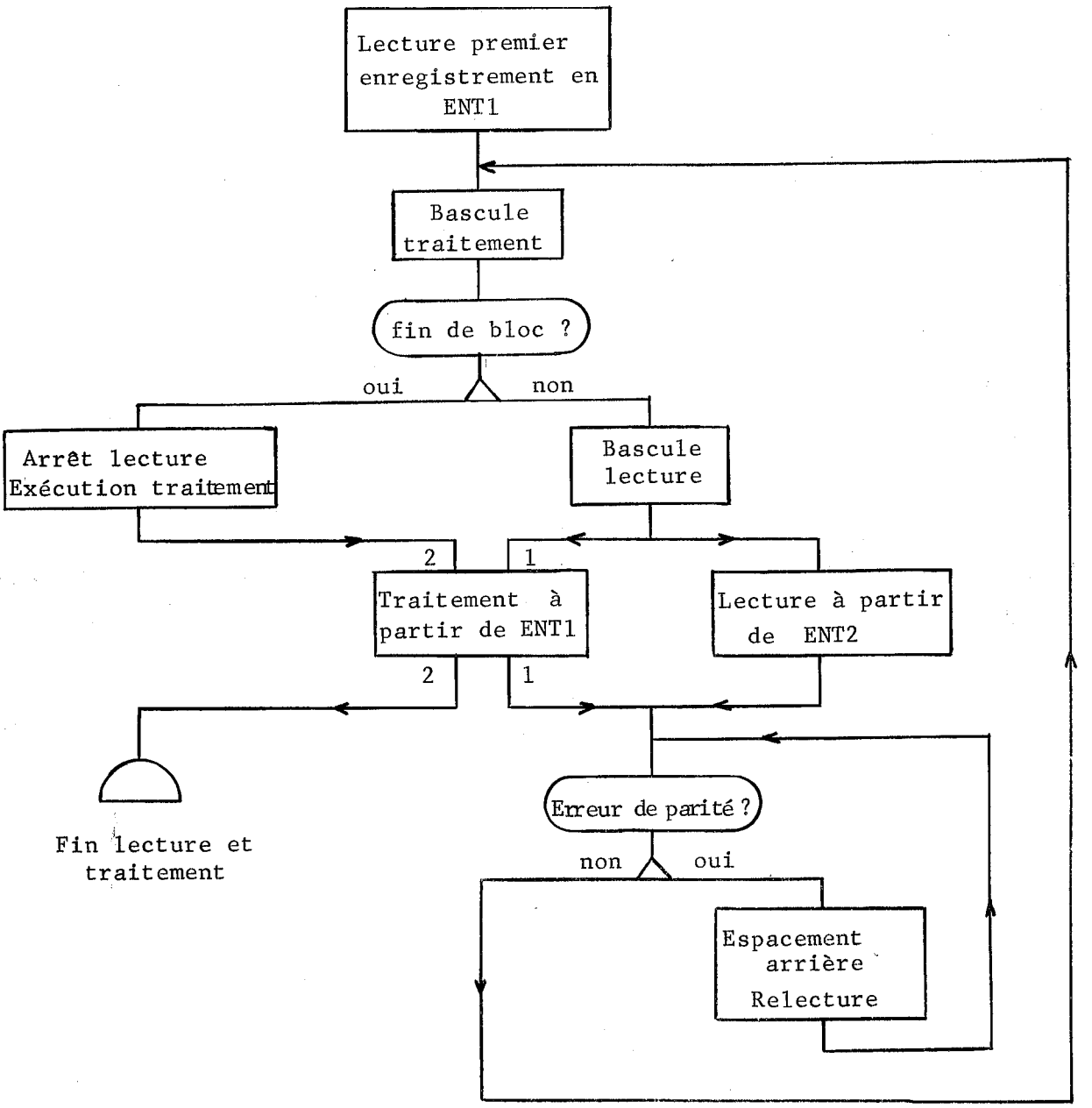
L'organisation des buffers est du type 2 . Le programme de gestion peut être appelé en un point quelconque du programme principal. Il est aussi utilisé pour remplir une fonction d'entrée (ou de sortie) et un traitement interne : cas de la sélection d'articles dans un dictionnaire.

Un certain nombre de tests sont prévus pour valider les transmissions : erreur de parité, fin de bande, enregistrements bruits, etc ...

Remarques :

- Ce programme n'envisage pas l'utilisation des disques comme moyens de stockage de l'information.

- Ce programme nous a permis d'étudier les techniques d'utilisation de la simultanéité qui ont été appliquées sous le système IBSYS 7090 [18] . Sur IBM 7044, le problème a été repris et traité d'une manière analogue, sous le système IBJOB 7044.



ORGANISATION GENERALE D'UN PROGRAMME LECTURE - TRAITEMENT EN SIMULTANEITE

CEUX AUXQUELS DU BONHEUR ET DE L'AFFECTION SONT DONNES Y SONT RAREMENT SENSIBLES A CAUSE D'UN AMOUR-PROPRE EXCESSIF .

ELLE T'EN DONNE .

J'Y SUIS ALLE .

JE M'ETAIS BRULEE .

NOUS L'ATTACHONS .

ILS S'ETAIENT HABILLES .

METS LE ATTACHE .

SUIS JE ALLE ASSEZ DROIT .

IL N'EN CUEILLE PAS .

C'EST LUI , S'IL DISSOLT ...

IL SAIT CE QU'ELLE CHANTE .

J'AI DONNE JUSQU'A MON EPOUSE .

QU'IL S'HABILLE .

TU ME DONNES LA SIENNE .

VOUS L'ATTACHEZ AU NOTRE .

DE QUOI SOUFFRAIT IL .

A QUI DONNAIT IL DES ORDRES .

CET AMATEUR D'EAU-DE-VIE N'EST PAS UN EQUILIBRISTE .

SON AVANT-DERNIERE CANDIDATE SAVAIT PATINER .

LE BEAU ET VIEIL HABILLEUR ETAIT SOURD-MUET .

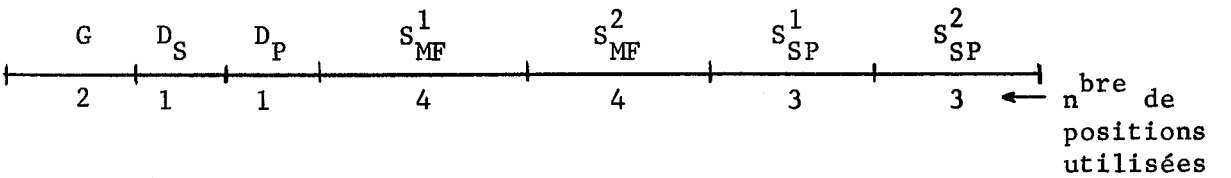
IL NE SAIT CE QU'IL VA METTRE .

A N N E X E 3

PROGRAMME DE SYNTHESE MORPHOLOGIQUE

I - FORME MACHINE DU CODE MORPHOLOGIQUE

1 - Substantif



G genre

D_S défectivité singulier

D_P défectivité pluriel

S^1_{MF} premier Système de terminaisons masculin - féminin ; il est nul si le mot n'est pas composé ou si $G \neq 3$.

S^2_{MF} deuxième Système de terminaisons masculin - féminin ; nul si $G \neq 3$

S^1_{SP} premier Système de terminaisons singulier - pluriel ; nul si le mot n'est pas composé ou si $D_S \cup D_P = 1$.

S^2_{SP} deuxième Système de terminaisons singulier - pluriel ; nul si $D_S \cup D_P = 1$.

E élision

CT contraction

N nombre

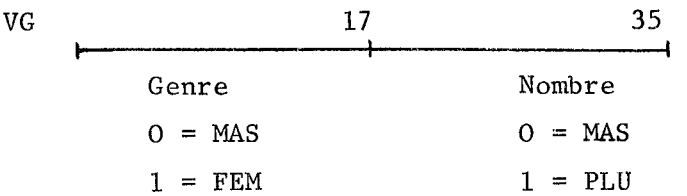
Remarque

Pour les différentes classes que nous venons de voir le code morphologique comporte trois caractères. Pour les autres classes soit il n'existe pas de code morphologique (adverbe), soit il est implicitement donné par l'ordre des bases (classes fermées).

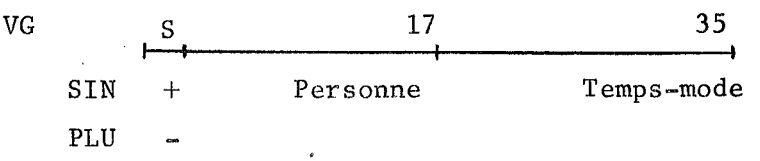
II - DISPOSITION DES VARIABLES GRAMMATICALES DANS LE PSEUDO-TEXTE

Nous avons vu que les variables grammaticales occupent la quatrième mémoire du syntagme élémentaire (chapitre VI, II). Le nombre et le nom des variables grammaticales changent selon la catégorie syntaxique.

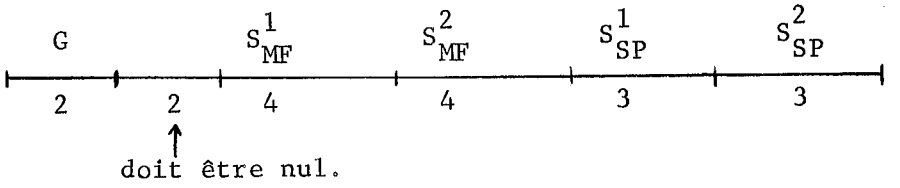
1 - Substantifs - Adjectifs qualificatifs et articles



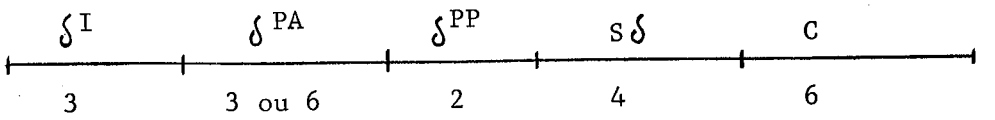
2 - Verbe (temps simples)



2 - Adjectifs qualificatifs, interrogatifs et indéfinis



3 - Verbe



δ^I désinence infinitif

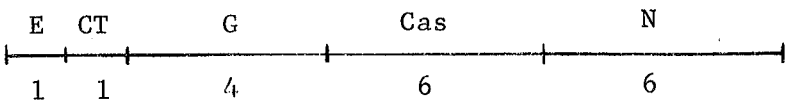
δ^{PA} désinence participe passé ; pour les verbes ABSOUDRE et DISSOUDRE, $\delta^{PA} = 70_8$ et occupe six positions (il n'y a pas d'ambiguïté car δ^I n'existe pas pour la même base).

δ^{PR} désinence participe présent.

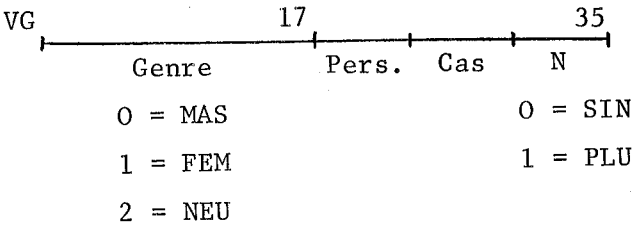
S δ numéro du système de désinences associé ; si S δ = 0 on a aussi C = 0 .

C numéro de catégorie.

4 - Pronoms interrogatifs et indéfinis



3 - Pronoms personnels faibles



Remarques :

- Dans certains cas cette mémoire est vide car il n'y a pas de variables grammaticales : infinitif, participe présent, etc...

- Nous n'avons pas donné tous les types de mémoire VG mais seulement les principaux.

VU,

Grenoble, le

Le Président de la Thèse

VU,

Grenoble, le

Le Doyen de la Faculté des Sciences

VU et permis d'imprimer,

Le Recteur de l'Académie de Grenoble