UNIVERSITE DE NANCY I

UER DE MATHEMATIQUES

RECONNAISSANCE DES FONCTIONS PRIMAIRES DE LA PHRASE ANGLAISE

THESE DE 3e CYCLE

présentée par

Jean-Bernard JOUIN

le 27 octobre 1973

JURY : président M. C. PAIR

examinateurs Mle H. NAÏS
M. G. BOURQUIN
M. J.C. DERNIAME

UNIVERSITE DE NANCY I

UER DE MATHEMATIQUES

RECONNAISSANCE DES FONCTIONS PRIMAIRES DE LA PHRASE ANGLAISE

THESE DE 3e CYCLE

présentée par

Jean-Bernard JOUIN

le 27 octobre 1973

JURY : président M. C. PAIR

examinateurs Mle H. NAÏS
M. G. BOURQUIN
M. J.C. DERNIAME

A Michèle, ma femme.

Je tiens à remercier :

- Monsieur PAIR, Directeur de l'I.U.C.A. qui a dirigé mon travail et m'a prodigué ses conseils notamment dans la partie de formalisation que ce travail comportait.

- Mademoiselle NAIS, Directrice du C.R.A.L. qui m'a accueilli dans son laboratoire, familiarisé avec les problèmes linguistiques et me fait aujourd'hui l'honneur de participer au jury.

- Monsieur BOURQUIN qui a mis au point avec la Section de Traduction Automatique qu'il dirige, toute l'analyse linguistique du problème. Cette section se compose de Mesdames Bourquin, Beer-Gabel, Euvrard et de Mademoiselle Lecompte et je remercie chacun de ses membres.

- Monsieur DERNIAME qui a accepté de bien vouloir faire partie du jury.

- Mesdames Thouvenot, Thiery et Mademoiselle Barraux qui se sont chargées de la perforation des programmes et des jeux d'essais.

- Le service d'exploitation de l'I.U.C.A. qui a assuré le passage des programmes et les ingénieurs-système qui m'ont aidé à résoudre quelques problèmes de software.

- Mesdames Humbert et Thouvenot, enfin, qui se sont occupées de la réalisation matérielle de cette thèse.

TABLE DES MATIERES

TABLE DES MATIERES

ANNEXES

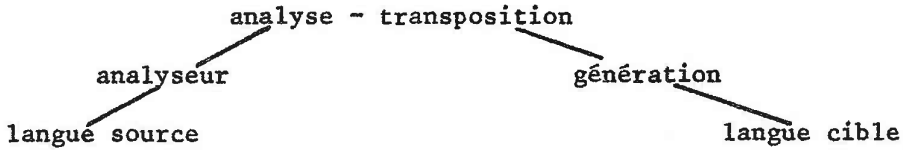
PREAMBULE

INTRODUCTION

P R E A M B U L E

Ce travail s'inscrit dans le plan général de recherche de la section de Traduction Automatique du Centre de Recherches et d'Applications Linguistiques (C.R.A.L.) de Nancy.

En effet, si l'on reprend le schéma du traducteur automatique présenté par Monsieur Bourquin, on découvre que la première partie de ce schéma est constituée par un analyseur syntaxique. C'est un analyseur qui ne va pas tout à fait au fond des choses, ne fournissant pas les analyses les plus fines. Néanmoins, il répond aux critères minimaux nécessaires à la poursuite du travail vers la langue cible, à savoir les algorithmes de transposition et enfin la démarche de génération.



Nous avons découvert, au fur et à mesure de nos investigations, toute l'importance de l'analyseur dans la chaine de traduction, mais également l'importance du travail préparatoire à accomplir avant d'en venir à la réalisation de l'analyseur proprement dit.

0. INTRODUCTION

0.1. Présentation du problème

Les linguistes de la section de Traduction Automatique du C.R.A.L. à Nancy étudient depuis de nombreuses années un important volume de textes écrits en Anglais scientifique.

Avec l'aide d'un ordinateur, ils ont obtenu de ces textes, préalablement codés, divers instruments de travail tels que : listes de mots, de groupes verbaux ou nominaux, de contextes de tel ou tel mot particulier.

Leurs investigations ont fait apparaître clairement la possibilité d'une analyse automatique de la phrase anglaise.

Aidés en cela par les outils fournis par l'ordinateur, ils ont pu dégager un ensemble d'algorithmes de reconnaissance des structures de surface de la phrase.

Dans un deuxième temps et en suivant une démarche analogue, ils ont mis au point une méthode de reconnaissance des fonctions grammaticales pour ces diverses structures.

C'est à l'exploitation sur ordinateur de ces algorithmes et au test de ces différentes hypothèses que le C.R.A.L. nous a demandé de réaliser à partir d'octobre 1971. Nous nous sommes fixé comme but de parvenir à la reconnaissance des fonctions sujet et objet en ayant soin d'éliminer au maximum les cas non prévus.

Nous appellerons par commodité ces fonctions : fonctions primaires de la phrase.

0.2. Explications succintes et buts

Par fonctions primaires nous entendons, en Anglais comme en Français, les fonctions syntaxiques de base que sont la fonction sujet et la fonction objet. C'est autour du verbe et à partir de ces fonctions primaires que s'organise la proposition et, par là, la phrase.

La méthode que nous avons employée ici s'apparente à celle utilisée par le traducteur humain, lorsqu'avant de traduire une phrase, il commence par l'analyser. Par conséquent, nous suivrons une démarche allant des mots aux groupes, des groupes aux syntagmes, pour parvenir aux propositions et aux fonctions des syntagmes dans celles-ci.

Pour déterminer la nature grammaticale d'un mot nouveau (non encor

connu de lui) l'être humain utilise un dictionnaire. Pour le calculateur tous les mots sont nouveaux : cette détermination requiert donc la consultation d'un dictionnaire stocké en mémoire de masse. Nous découvrirons à ce niveau tout l'intérêt que peut présenter le problème du stockage et de la consultation de ces informations.

De nombreux mots présentent plusieurs analyses possibles, mais dans un contexte donné l'ambiguïté disparaît. Il nous aura donc fallu déterminer les mécanismes et les lois de ces levées d'ambiguïtés.

Il existe enfin dans la phrase des suites de mots qui jouent le rôle de mots isolés ou de groupes plus simples. Donnons-en pour exemple les substantifs qualifiés par un épithète, les temps composés des verbes, les formes passives etc...

Il nous a donc paru judicieux, dans notre optique de recherche des fonctions, d'élaborer les règles de ces regroupements. Ainsi serons nous amenés à parler de regroupement verbal ou nominal. La philosophie générale de cette recherche, en ce qui concerne la prise en compte d'une phrase consiste à la considérer comme une suite d'éléments, c'est-à-dire d'analyses de mots devenant successivement : des mots analysés (par les levées d'ambiguïtés), des éléments de groupes (par les regroupements), des éléments de groupes fonctionnels (par les attributions de fonctions). Schématiquement le texte subira successivement les transformations suivantes :

. Reconnaissance des mots
. Affectation des analyses
. Construction du groupe verbal
.. analyse de ses constituants
.. constitution du groupe
. Construction du groupe nominal
.. levées d'ambiguïté sur ses termes
.. constitution du groupe.
. Exploration des contextes des différents groupes verbaux selon les types de propositions auxquelles ils appartiennent en vue de trouver leurs sujets et leurs objets.

Chapitre I

LE DICTIONNAIRE

I.1. Nécessité d'un dictionnaire

Dire que rien ne peut être fait en linguistique sans dictionnaire est une évidence. Qu'il s'agisse en effet de codage de textes, d'études grammaticales ou de traduction, le dictionnaire s'avère indispensable.

Citons, à ce propos, KAY de la RAND-CØ (1967) [10] : "La linguistique a un besoin essentiel de données. Avec un ordinateur nous pouvons accumuler de grands fichiers contenant des textes variés : chercher si tel phénomène particulier s'y trouve et utiliser cela pour tester nos hypothèses plus ou moins plausibles... Nous avons commencé à constituer des fichiers de textes anglais ou russes avec ou sans informations subordonnées aux structures grammaticales ou sémantiques".

C'est ce même souci qui a poussé la Traduction Automatique du C.R.A.L. à créer un dictionnaire grammatical de formes anglaises, le plus complet possible. De nombreuses formes sont ambigues, c'est-à-dire se rencontrant dans des textes avec des analyses grammaticales diverses. Notre préoccupation a moins porté sur le nombre de formes de notre dictionnaire que sur la certitude d'avoir recensé toutes les analyses possibles de ces formes. On comprend alors aisément pourquoi ces formes doivent être tirées d'un volume de textes, d'un corpus important. En effet, plus une forme aura été rencontrée souvent et plus on approchera de cette certitude d'avoir rencontré toutes ses analyses. Par conséquent, la probabilité de découvrir ultérieurement des analyses nouvelles sera extrêmement faible. Nous disposons actuellement d'un vocabulaire comportant 10.000 formes extraites d'un corpus d'environ 150.000 mots.

Les mots les plus fréquents peuvent avoir 200 ou 300 occurrences, les plus rares moins de 10. Ces mots représentent globalement aux alentours de 17000 analyses. Le degré d'ambiguïté moyen (c'est-à-dire le nombre moyen d'analyses par mot) est donc d'environ 1,7. Pour certains mots on a rencontré des degrés compris entre 1 et 5.

Le vocabulaire appartient à l'Anglais scientifique puisque le corpus a été constitué par 71 articles extraits de revues portant sur la biologie, la cartographie, les mathématiques et la géographie.

Enfin, le nombre de 10.000 formes différentes dépasse de beaucoup les 4.000 utilisées par l'Anglais courant et nous paraît donc suffisant.

I.2. Extensions

C'est plutôt la taille du corpus initial qui nous a semblé insuffisante. C'est pourquoi nous avons entrepris récemment la saisie d'une série d'articles tirés du magazine bilingue 'IMPACT'. Cette saisie respecte les règles de préédition minimale, que nous exposons plus loin, et fait partie de la création d'un 'Trésor de la langue anglaise' qui comportera environ 1 million de mots.

Nous disposons de programmes permettant l'édition, à la demande des linguistes, de toutes sortes de listes de mots avec leurs fréquences, références et contextes plus ou moins larges à volonté. Ainsi, les chercheurs disposeront d'un instrument de travail de base tant pour parfaire, compléter, étendre le dictionnaire que pour améliorer leur connaissance des mécanismes de la syntaxe.

I.3. Règles d'une préédition minimale

Puisque l'on a rejeté dans un avenir assez lointain la possibilité de traduire la langue de façon automatique à partir de la parole, il nous faut nous concentrer sur un système de saisie et de reconnaissance du texte écrit.

Par définition, le mot est constitué d'une suite de lettres ou de chiffres précédée d'un espace et suivie d'une ponctuation ou d'un espace. On découvre immédiatement que ce mode de saisie couramment utilisé en typographie ne peut nous convenir. En effet, les chiffres décimaux peuvent comporter des points ou des virgules et constituent donc, au sens de notre première définition, des mots ambigus. Si l'on veut éviter ce genre d'ambiguité il est donc nécessaire d'imposer au texte, pour son entrée en machine, un format bien défini, en d'autres termes, un minimum de préédition. La préédition est encore imposée par 3 contraintes inhérentes aux ordinateurs actuels :

- Impossibilité de distinguer les majuscules
- Absence de certains signes spéciaux.
- Impossibilité de reconnaître, à priori, les expressions intraduisibles du type "en français dans le texte" et d'une façon générale, impossibilité de reconnaître et de transcrire les différents types de caractères qui sautent aux yeux de n'importe quel lecteur :

- titres, caractères gras, italiques, renvois, citations, etc... La préédition minimale est donc constituée des règles suivantes :

- On remplace la majuscule par la lettre normale précédée d'une astérisque (*) mais seulement dans le cas des noms propres et non pour les débuts de phrase.

- Les titres commencent par un symbole spécial (@) en raison de leur caractère fréquemment antigrammatical.

- Les ponctuations deviennent des mots à part entière par leur mise entre espaces. Ainsi disparaissent les ambiguïtés qu'elles auraient pu facilement créer.

I.4. Construction et édition du dictionnaire

Pour la construction du dictionnaire, nous avons procédé en deux étapes : tout d'abord l'élaboration d'un noyau d'environ 700 formes, puis son élargissement progressif.

Pour construire le noyau, nous avons choisi un ensemble de 33 phrases particulièrement représentatives dans différents articles de notre corpus.

Manuellement, en utilisant les listings de contextes, les linguistes ont affecté à ces formes, et avec un grand soin, le codage le plus complet possible.

Un premier test des programmes de traitement des textes a permis de donner à ce noyau sa forme définitive. Par la suite nous avons traité des textes nouveaux en utilisant ce noyau comme référence et les extensions du noyau ont de cette manière pu être faites d'une façon aussi précise que possible.

I.5. Passage dictionnaire

I.5.1. Définition

Par"passage dictionnaire" nous entendons : affectation à chaque mot d'un texte des codes grammaticaux trouvés sous sa rubrique dans le dictionnaire. On indiquera par exemple si le mot est un verbe, un substantif... s'il est au pluriel, au singulier etc...

I.5.2. Mode de codage employé pour le dictionnaire.

Nous avons décidé, afin que la mise à jour soit plus simple et compte tenu de la taille raisonnable du dictionnaire, d'attribuer à chaque mot autant d'enregistrements que ce mot présente de possibilités gramma-

ticales. Ces divers enregistrements (dont le nombre peut varier de 1 à 5) seront identifiés par un numéro d'ambiguïté figurant en première position de l'enregistrement.

Chaque enregistrement sera découpé en trois parties comportant des renseignements d'ordre :

. lexical : le mot
. grammatical
. syntaxique.

En conséquence, chaque enregistrement aura le format suivant :

1	2 26	27 32	33	34 53	54 59
numéro d'ambiguité	M Ø T	Code grammatical	Lexie	Catégories syntaxiques	Rections

Avant de développer les divers contenus de ces zones, il nous faut apporter quelques précisions au sujet de deux d'entre elles : ce sont lexie et rection. Nous appelons "lexie" un groupe de mots indissociables qui joue dans la phrase un rôle bien précis. Afin de faciliter la reconnaissance dans le texte de ces lexies, nous avons décidé de privilégier un mot pertinent ou terme central de la lexie. C'est cette possibilité du mot d'être pertinent que nous allons marquer dans la position 33. D'autre part, certains adjectifs verbes ou substantifs anglais ont la propriété d'être fréquemment suivis par des groupes nominaux introduits par une préposition spécifique. L'exemple le plus frappant est donné par le verbe TODEPEND qui introduit pratiquement toujours la préposition ON.

Les prépositions étant en nombre fini et toutes connues, nous avons décidé d'affecter à chacune d'elles un code spécifique. Nous pourrons ainsi marquer dans la zone rection le code de la préposition appelée.

. Catégories syntaxiques.

L'étude de notre corpus nous a permis de découvrir que certains mots avaient la possibilité d'introduire des constructions d'un type bien défini. Ainsi, un verbe transitif pourra introduire un groupe nominal ou une conjonctive.

Ces constructions ont été répertoriées et pour de nombreux mots on est capable d'attribuer une liste de constructions possibles introduites par eux. Ces mots recevront donc au dictionnaire une liste de codes syntaxiques correspondant chacune à une construction bien définie.

Ces constructions sont les suivantes :

Code	Nature
01	groupe adjectival
02	groupe nominal
03	conjonctive commençant par "THAT'
04	groupe prépositionnel
05	verbe à l'infinitif
06	groupe nominal suivi de infinitif sans 'TO'
07	groupe nominal suivi de groupe prépositionnel
08	deux groupes nominaux
10/30	groupe nominal suivi d'un infinitif avec 'TO'
11	verbe en ing
12	une forme interrogative
13	une construction impersonnelle + conjonctive avec THAT
14	une préposition
15	une construction impersonnelle suivie d'un infinitif
23	construction impersonnelle + adjectif + THAT
24	construction impersonnelle + adjectif + infinitif
34	préposition suivie de verbe en ing

Ainsi on pourra trouver dans l'une des catégories d'un mot, susceptible d'introduire une conjonctive commençant par THAT, la catégorie 03.

Remarquons que ce code a été choisi de telle sorte que tous les mots possédant des rections comportent au moins l'une des catégories syntaxiques : 04, 14, 34.

. Code grammatical :

C'est un code à 6 positions dans lequel la position 3 joue un rôle prépondérant : celui de donner au mot sa nature grammaticale brute : préposition, adverbe, substantif, etc...

Tableau des codes

1	2	3	4	5	6	Signification
		A				adjectif
		B				adverbe
		C				subordonnant
		D				coordonnant
		E				coordonnant
		F				cardinal
		G				préfixe
		I				introducteur
		K				pronom
		M				interrogatif
		P				préposition
		S				substantif
		U				'NØT'
		V				verbe
		X				auxiliaire
		Y				défectif
			'b'			genre, nombre indéterminés
			1			adjectifs : épithète
			2			adjectifs : attribut
			A...Z			codes spécifiques pour prépositions ou subordonnants
			S,P			nombre : introducteurs, pronoms, substantifs
			D,S,G			terminaison des verbes (ED, S, ING, EN...)
			N			
	1					adjectifs, adverbes : marques de comparatif
	1,2,3,4					auxiliaires : marque distinctive
	H					marque d'abréviation
U						groupe à un seul terme
b						nombre de termes indifférent
D						groupe à plusieurs termes
				A		verbes : voix active
				P		verbes : voix passive
				'b'		verbes : voix ambigüe ou absente

Comme on l'a vu, chaque préposition reçoit un code, bien défini et ce code se trouve dans les positions 3 et 4 du code grammatical.

Ces codes sont :

Code	prép.
PA	OF
PB	THAN
PC	TO
PD	AS
PE	SINCE
PF	FOR
PG	IN

Code	prép.
PH	WITH
PI	FROM
PJ	ON
PK	'S
PL	INTO
PM	AFTER
PN	AT
PO	BETWEEN
PT	ABOUT

Code	prép.
PQ	BY
PR	OVER

I.5.3. Diverses méthodes.

Nous proposons 2 méthodes permettant de résoudre le problème posé par la consultation du dictionnaire. Ces méthodes se définissent par le type d'accès que nous avons au dictionnaire : accès séquentiel ou accès direct. Le point de départ commun à ces deux méthodes est un éclatement du texte en une suite de mots référencés dans leur texte d'origine. Le format des enregistrements est le suivant :

REF	LG	MAJ	MOT

REF : référence du mot dans le texte (en général un indicatif de texte suivi d'un numéro d'ordre dans le texte).

LG : le nombre de caractères du mot.

MAJ : '*' si l'initiale du mot est une majuscule sinon un blanc.

Le résultat a atteindre est un fichier dans lequel les mots se trouvent codés dans l'ordre du texte avec pour chaque mot autant d'enregistrements que ce mot présente d'analyses. Les références sont donc complétées par un chiffre indiquant le n° de l'analyse.

REF	N°	LG	MAJ	MOT	CODAGE

N° : le n° de l'analyse du mot

CODAGE : le code affecté.

Ces deux fichiers sont utilisés en format fixe et la longueur des mots a été surdimensionnée à 25 caractères.

I.5.3.1. Méthode séquentielle.

Le dictionnaire étant trié par ordre alphabétique, il nous faut, pour traiter les mots séquentiellement, les mettre auparavant par ordre alphabétique. Enfin, il faut après l'affectation, restituer le texte dans son ordre original en le triant grâce aux références.

Cette méthode un peu lourde présente l'avantage d'autoriser le traitement d'un texte arbitrairement long. C'est cette méthode que nous avons utilisée exclusivement en raison de l'espace disque relativement restreint dont nous disposions pour nos exploitations.

Comme on le verra plus loin en raison des performances très raisonnables du processeur de tri les temps de traitement n'ont jamais été prohibitifs.

I.5.3.2. Accès direct.

Nous allons partager notre dictionnaire en plusieurs microglossaires sur lesquels nous aurons un accès direct.

Une façon assez intéressante de faire le partage consiste à partager selon la première lettre des mots. Nous dirons par exemple que le premier glossaire comporte les mots dont la première lettre est A ou B... le dernier, ceux qui commencent par les lettres W, X, Y, Z. Ainsi, lisant un mot dans le texte, un simple test sur sa première lettre nous indiquera le glossaire à consulter.

Le partage en plusieurs fichiers présente un gain de temps appréciable puisque l'accès à un fichier beaucoup moins volumineux est naturellement plus rapide.

I.5.4. Mots non affectés par le dictionnaire.

Il existe toute une série de mots qui ne reçoivent pas de code en provenance du dictionnaire. Il peut s'agir de mots inconnus, c'est-à-dire oubliés ou mal orthographiés et ils sont alors signalés dans une liste d'anomalies. Il peut s'agir également des ponctuations ou des mots qui sont alors codés automatiquement (ex. nombres cardinaux et ordinaux).

I.5.5. Calcul des temps moyens d'affectation.

I.5.5.1. Accès direct.

Nous utilisons un dictionnaire éclaté en 3 fichiers d'accès direct.

1°/ Lettres A à H : DICA 310 enregistrements
2°/ Lettres I à Q : DICI 232 enregistrements
3°/ Lettres R à Z : DIER 258 enregistrements

Temps de création : (en minutes)

Création	UC	WAIT	TOTAL	TAILLE
DICA	0.01	0.04	0.05	310
DICI	0.00	0.04	0.04	232
DICR	0.00	0.04	0.04	258
TOTAL	0.01	0.12	0.19	

Avant toute exploitation, les fichiers sont à restaurer.
Temps de restauration (bande à disque) en minutes.

Restauration	UC	WAIT	TOTAL
DICA	0.01	0.11	0.12
DICI	0.00	0.07	0.07
DICR	0.00	0.07	0.07
TOTAL	0.01	0.25	0.26

Pour les exploitations ultérieures, ce seront donc les temps de restauration qui seront à prendre en compte.

La méthode d'accès direct a été testée de deux manières distinctes sur l'ordinateur C.I.I. 10.070 de l'I.U.C.A. sous le système SIRIS 7.

1°) Sur un texte réel comportant 3200 mots par les affectations successives de 300, 1000, 2000 et de la totalité des mots.

Nous avons relevé les temps suivants :

(y compris le temps de restauration) en minutes

Nombre de mots	NOmbre d'accès disques	UC	WAIT	TOTAL	temps moyen par accès ms.
300	460	0.28	0.82	1.10	125 ms
1000	1650	1.18	1.75	2.93	113 ms
2000	3400	2.19	3.08	5.27	93 ms
3200	5100	3.80	3.44	6.64	78 ms

2°) Par simulation.

Nous avons choisis 3 mots, 1 dans chaque index. Les 3 mots sont ambigus et nécessitent respectivement 3, 3 et 4 accès disque pour leur affectation. Nous les avons affectés au moyen d'une boucle effectuée successivement 500, 1000 et 1500 fois, ce qui nous a conduit respectivement à 1.666, 3.333 et 5.000 accès-disque ou consultations des 3 index. Les temps relevés (y compris restauration) sont :

Nombre d'accès	UC (min)	WAIT (min)	TOTAL (min)	Temps moyen ms.
1.666	0.22	2.58	2.80	100.8 ms
3.333	0.41	4.20	4.61	83 ms
5.000	0.58	5.06	5.64	68 ms

I.5.5.2. Accès séquentiel.

Dans cette méthode, on ne verra plus apparaître les temps de restauration du dictionnaire puisque tout le traitement est fait sur bande magnétique.

Nous avons relevé les temps suivants :

Livre	Nombre de mots	Nombre d'analyses	Ier TRI		Affectation		2e TRI		Temps total	Temps moyen
			UC	WAIT	UC	WAIT	UC	WAIT		
1	4041	6058	0.29	0.20	0.60	0.20	0.35	0.30	1.94m	19,4 ms
2	1553	2174	0.11	0.14	0.32	0.09	0.15	0.19	1.00m	27,3 ms
3	1203	1924	0.08	0.19	0.24	0.13	0.12	0.24	1.00m	31,6 ms
4	1216	1824	0.08	0.22	0.26	0.19	0.13	0.34	1.22m	40,6 ms
5	1826	2556	0.13	0.26	0.37	0.22	0.19	0.40	1.57m	36,2 ms
6	1243	1988	0.09	0.29	0.31	0.26	0.15	0.44	1.54 m	46,1 ms
7	994	1471	0.07	0.33	0.21	0.27	0.10	0.50	1.48m	59,2 ms
8	1395	1907	0.10	0.37	0.34	0.28	0.16	0.54	1.78m	56,2 ms

I.5.6. Conclusion :

La méthode séquentielle possède un avantage non négligeable : c'est de ne coder qu'une fois les mots identiques. L'expérience nous a montré que souvent 10 à 15 % des mots forment plus de 50 % de l'ensemble du texte : auxiliaires, articles, etc...

Les temps que nous avons indiqués nous montrent combien cette méthode quoique lourde, présente d'avantages.

Si maintenant nous comparons avec les temps moyens trouvés pour l'affectation d'une analyse à un mot, nn peut faire les remarques suivantes :

Dans les deux méthodes le temps moyen est inversement proportionnel au nombre total de mots à affecter ; cette première remarque entre bien dans la logique des choses puisque les temps de restauration, ouverture de fichiers etc... interviennent toujours le même nombre de fois indépendamment du nombre de mots à traiter.

Pour toutes les valeurs du nombre de mots affectés, la méthode par accès séquentiel s'avère toujours plus performante.

Chapitre II

LE TRAITEMENT DES LEXIES, PREFIXES, RECTIONS, GROUPES A TRAIT D'UNION

II.1. BUTS

Parvenus à ce point du traitement du texte, nous disposons des mots affectés des informations contenues dans notre dictionnaire.

Nous avons à présent à reconnaître et regrouper "dans le texte" les lexies, mots à préfixes et groupes à trait d'union. Enfin, pour les mots suivis de prépositions bien définies, que nous appelons rections, nous allons tester si ces rections sont effectivement présentes et marquer dans le code des mots eux-mêmes cette réalisation.

II.2. FORMATION DES LEXIES.

II.2.1. Définition :

On appelle lexie une suite figée de mots qui s'est formée par un phénomène que les linguistes appellent lexicalisation. La lexicalisation a pour effet de donner à certaines suites de mots des contraintes tellement fortes que ces mots deviennent indissociables.

La plupart du temps les lexies jouent dans la phrase le rôle de mots invariables : prépositions ou adverbes.

On peut citer en exemple l'expression : AND SO ON qui trouve bien son équivalent français dans ET AINSI DE SUITE. Comme de nombreux mots, les lexies peuvent présenter plusieurs codes grammaticaux : ainsi UP TØ qui peut, selon le contexte, être une préposition ou un adverbe.

La fréquence en anglais de ces expressions nous a incité à leur accorder un traitement particulier.

II.2.2. Répertoire des lexies.

Plus peut-être que d'autres, la langue anglaise présente de nombreuses lexies. C'est pourquoi il nous a paru judicieux d'en dresser une liste, un dictionnaire qui soit aisément extensible. Pour ne pas avoir à consulter ce dictionnaire pour chaque mot de la phrase, on va déterminer de manière unique, pour chaque lexie, un terme central que nous appellerons "mot pertinent". Dans le dictionnaire général des mots cette faculté d'être mot pertinent d'une ou plusieurs lexies sera donc marquée.

Afin de ne pas allonger la liste des lexies figurant sous un même terme central, nous choisirons pour termes centraux des mots assez peu fréquents de la langue.

Le terme central fournira au dictionnaire des lexies, avec son numéro d'ordre, la "clé" d'un enregistrement où l'on trouvera encore :

- La lexie complète avec son codage
- Un pointeur d'ambiguïté
- Le nombre de mots et la place du terme central dans la lexie.

Pointeur	terme central				lexie....	Code	
1	SET	2	1	01	SET FORTH	VDP	00
1	SET	2	1	02	SET UP	VDP	02
2	SET	2	1	02	SET UP	VDP	00

La rencontre de SET dans un texte nous permettra de tenter un regroupement avec le mot suivant et de vérifier si l'une ou l'autre des lexies réalisables avec ce mot est effectivement réalisée.

II.2.3. Conclusion :

Nous avons jusqu'à présent retenu environ 200 lexies différentes. Le programme de formation des lexies "dans le texte" est, bien entendu, accompagné d'une procédure de mise à jour, d'un emploi facile, du fichier des lexies. Cette procédure autorise de façon simple l'insertion, le remplacement ou la suppression d'une lexie quelconque. Actuellement pour un dictionnaire d'environ 1.600 mots, nous avons dégagé un peu plus de 200 lexies s'appuyant sur : 110 mots centraux distincts.

II.3. PREFIXES, TRAITS D'UNION, RECTIONS.

II.3.1. Préfixes, traits d'union.

On a vu que dans le dictionnaire les préfixes sont signalés par la lettre 'G' dans la 3e position de leur code grammatical. En anglais, les préfixes ne forment pas toujours avec le mot qu'ils qualifient un mot unique. Si en français on trouve le terme 'acide isobutyrique' par contre en anglais on trouvera 'iso butyric acid'. Ainsi le préfixe reste séparé de son substantif de rattachement.

Pour ne pas avoir à faire figurer tous ces termes dans le dic-

tionnaire des lexies, nous nous sommes donnés la règle suivante :

. le préfixe sera regroupé dans le texte avec le mot qui le suit.

Cette règle est également valable pour deux mots séparés par un trait d'union. L'unité résultant du regroupement reçoit pour code :

- Celui du dernier terme dans tous les cas sauf si ce dernier terme est un verbe en 'ed' ou 'ing'. C'est alors le code adjectif qui est attribué.

Exemple : amino (G) acid (ss) ⟶ amino acid (ss)

multi (G) - lens (SP) ⟶ multi-lens (SP)

par contre : diesel (A/SS) - engined (Vd) ⟶ diesel-engined (A)

II.3.2. Rections :

On a vu que certains adjectifs, substantifs ou verbes peuvent appeler des prépositions spécifiques. Si l'hypothèse est réalisée et que la préposition soit effectivement présente, on le signale dans son code. Ainsi dans la suite : ... DEPENDS ON... codée Vga... Pj la préposition devient PJV, marquant ainsi son caractère de rection verbale.

Cette information constitue un pas déjà très important vers la recherche des circonstants. En effet la préposition est forcément suivie d'un groupe nominal ; et si cette préposition est une rection verbale, le groupe nominal est presque sûrement le circonstant du verbe.

Chapitre III

LA CHAINE VERBALE

III.1. LEVEE DES AMBIGUÏTES VERBALES

III.1.1. Définition :

On appelle mot ambigu verbal ou ambiguïté verbale tout mot ambigu dont l'un au moins des codes est un code verbal.

III.1.2. Liste des ambiguïtés verbales

Sigle	Codes	Contenu
V20	VSA / SP	ambigu V à la 3e personne et subst. pluriel
V21	VOA / SS	ambigu verbe et substantif singulier
V22	VRA / VRP / A / SS	verbe ambigu actif-passif ou adj. ou substantif
V23	VRa / VRP / SS	verbe ambigu actif-passif ou adj. ou substantif
V25	V-A / A	verbe actif ou adjectif
V26	Voa / P	verbe actif ou préposition
V27	Voa / A / SS	verbe actif adjectif ou substantif
V28	Voa / A / B	verbe actif adjectif ou adverbe
V29	VGa / P	verbe en ing ou préposition
V30	VGa / C	verbe en ing ou conjonction
V31	Vga / A1	verbe en ing ou adjectif
V32	Vga / SS	verbe en ing ou substantif
V33	Vga / A1 / SS / P	verbe ou adj. ou subst. ou préposition
V34	Vga / A1 / SS	verbe ou adj. ou subst. ou préposition
V35	V-e / V-P	ambigu actif / passif
V36	Vda / Vdp / C	ambigu actif/passif ou conjonction
V37	V-p / A1	verbe passif ou adjectif
V38	Vda / A	verbe actif ou adjectif
V 39	Vdp / SS	verbe passif ou substantif
V40	Va/V-p / A	verbe actif ou passif ou adjectif
V41	X / V	auxiliaire ou verbe
VV42	Y / V / SS	défectif verbe ou substantif

III.1.3. Classification et traitement.

Nous relevons dans le dictionnaire des mots trois grandes classes d'ambiguïtés verbales.

1) L'un des codes est verbal, les autres ne le sont pas. On trouvera dans cette classe tous les mots pouvant être : verbe ou substantif, verbe ou adjectif, verbe ou pronom etc...

2) L'un des codes est verbal, un second auxiliaire. Nous trouvons dans cette catégorie tous les auxiliaires et défectifs de la langue anglaise tels que HAVE, BE, MUST, DO etc... qui selon le contexte ont dans la phrase un rôle d'auxiliaire ou simplement une fonction verbale.

3) Deux codes au moins sont verbaux.

Nous rangeons ici la plupart des verbes en -ed qui peuvent être soit verbes conjugués au prétérit soit auxiliés.

Si l'on désigne par degré d'ambiguïté d'un mot le nombre d'entrées dont on dispose pour ce mot dans le dictionnaire, on remarquera que ce degré est compris entre 2 et 4 pour la première catégorie, qu'il vaut toujours 2 pour la seconde et enfin qu'il est compris entre 2 et 5 pour la troisième. Notre programme de levée d'ambiguïté utilise, stockée en mémoire centrale une table de toutes les ambiguïtés verbales connues. Chaque ligne de cette table donne accès à un traitement spécifique qui sera mis en oeuvre lorsque l'ambiguïté aura été trouvée dans le texte.

Pour ce faire, la phrase est entièrement mise en mémoire et, par la suite, chacun de ses mots est considéré comme un élément d'un tableau. Chaque mot de la phrase reçoit donc un ensemble d'indices consécutifs (un indice par code grammatical). Ces indices sont les paramètres des traitements de levée d'ambiguïté.

L'examen des prédécesseurs et successeurs du mot ambigu dans la phrase doit permettre de résoudre l'ambiguïté. Le résultat du traitement est alors l'indice du code grammatical choisi. Si le contexte s'avère cependant insuffisant pour lever l'ambiguïté, c'est la valeur 0 qui est rendue par le traitement.

Aucun des traitements prévus n'exclut la sortie ambigue, c'est-à-dire le résultat 0. Les numéros des analyses écartées par une

levée d'ambiguïté effective sont mis à zéro et ces numéros sont réorganisés lorsque toute la phrase a été traitée. Enfin la phrase épurée des analyses superflues est recopiée dans le fichier résultat. A tous les niveaux de l'analyse syntaxique nous aurons encore à effectuer des levées d'ambiguïtés. Le canevas que nous venons de présenter pour les verbes restera bien entendu valable et les trois traitements successifs (reconnaissance de l'ambiguïté, examen du contexte et résolution dans un sous-programme spécial, épuration et recopie des mots analysés) seront présents dans toute levée d'ambiguïté. L'exemple que nous donnons ci-dessous fait apparaître l'état d'une phrase avant et après le passage d'un programme de levées d'ambiguïtés verbales.

III.1.4. Exemples :

Dans cette phrase on a pu lever trois ambiguïtés verbales grâce à un contexte significatif. On voit aussi apparaître par trois fois les numéros des analyses superflues remis à 0 et les valeurs rendues par les sous-programmes mis en oeuvre.

MOTS	INDICES	N°	ANALYSES	INDICE RESULTAT	N°	ANALYSES-RESULTAT
SEDIMENTARY	1	1	A1	1	1	A1
ROCKS	2	1	SP		1	SP
MAY	3	1	YM		1	YM
BE	4	1	1XO		1	1XO
	5	2	1VOA	4	0	
GROUPED	6	1	VDP		1	VDP
IN	7	1	PG		1	PG
TWO	8	1	IP		1	IP
	9	2	KP	0	2	KP
MAIN	10	1	A		1	A
CLASSES	11	1	SP		1	SP
:	12	1			1	
CONSOLIDATED	13	1	VDA		0	
	14	2	VDP	15	0	
	15	3	A1		1	A1
AND	16	1	E1		1	E1
UNCONSOLI-DATED	17	1	A1		1	A1

III.1.5. Règles des levées d'ambiguïtés.

III.1.5.LListe des codes synthétiques utilisés

CODE 1, CODE 2, ADV 1... ADV 6 représentent des listes de codes grammaticaux possibles. Ces termes seront employés dans les schémas (III.1.5.) des levées d'ambiguïtés verbales.

Ces codes d'nne même liste ont pour particularité d'être attribués à des mots qui, dans un contexte bien précis, fonctionnent de la même façon. Il est alors plus simple de se demander si un mot est du type ADV 1 que d'énumérer une longue série de tests sur les codes de cette même liste.

Les listes CODE ... concernent des verbes.
Les listes ADV ... concernent des adverbes.

CODE 1	CODE 2
Vna, Vta, Vta, Vda	Vnp, Vdp, Vtp, Vtp
Vna/Vnp, Vda/Vdp	Vna/Vnp, Vda/Vdp
Vta/Vtp, Vna/Vnp	Vta/Vtp, Vta/Vtp
Vna/Al, Vda/Al	Vnp/Al, Vdp/Al
Vta/Al	Vtp/Al
Vna/Vnp.Al	Vna/Vnp/Al
Vda/Vdp/Al	Vda/Vdp/Al
Vta/Vtp/Al	Vta/Vtp/Al
Vda/Ss	
Vda/Al/Ss	
Vda/Al/Ss/B	

ADV1	ADV2	ADV3
B, A/B, A1/B1, 1A/1B I/B, I/K/B I/A2/B, I/K/A/B Is/Ks/D2/B Pe/Be/Ce Pd/Cd/2B,C/B	A/B, A1/B1 L1A/1B, I/B I/K/B, I/A2/B I/K/A/B I/K/Ss/B Vda/Vdp/A/Ss/B G/I/Ss/B P/A/B	B, I/B, A/B C/B, Pa/Cd/2B Pe/Ce/Be, A/S1/B I/K/B, Is/Ks/D2/B Bda/Vdp/A/SS/B

ADV4	ADV5	ADV6
B, A/B, I/B I/K/B, I/K/A/B Is/Ks/D2/B	A/Ss/B I/K/Ss/B G/I/Ss/B Vda/VdP/A/Ss/B	P/B, Pe/Ce/Be P/C/B, P/A/B Pd/Cd/2B, CB B2, hB

5.2.

III.1. SCHEMAS DES PROCESSUS DE LEVEES D'AMBIGUITES.

Les algorithmes n'utilisent que le contexte du mot ambigu. Nous dirons que les ambiguïtés sont levées par des impossibilités de succession. De plus la longueur du contexte à explorer ne dépasse jamais 3 mots en amont ou en aval du mot ambigu.

On peut donc représenter les règles de levée d'ambiguïté par un tableau à 9 colonnes.

Colonne 1 : un codage de l'ambiguïté à lever

Colonnes 2,3,4 : codage des prédécesseurs de rang -3, -2, -1 du mot ambigu.

Colonne 5 : l'ambiguïté à traiter

Colonnes 6,7,8 : codage des successeurs aux rangs +1, +2, +3 du mot ambigu.

Colonne 9 : le résultat donné par l'algorithme.

Le mode de l'écriture des codages appelle quelques explications :

/ désigne un mot ambigu (V/S : ambigu verbe, substantif)

, indique un choix entre plusieurs codages

désigne la négation d'un code (¬ P : tout code sauf P)

PONCT terme général désignant une ponctuation

(codage) : la présence d'un mot à sauter (d'un code non obligatoire)

ainsi | -1 / (A),B | désigne à la fois | -2 / B | -1 / A | et | -1 / B |

une colonne sans code indique que le successeur ou le prédécesseur se trouvant là n'est pas testé par l'algorithme :

'mot' indique un mot spécifique à tester.

Cod1, adv1... le mot à tester appartient à l'une des listes générales de noms cod1, adv1...

Résolution des ambiguïtés : verbales

PREDECESSEURS			AMBIGUITE	SUCCESSEURS			RESULTAT
-3	-2	-1		+1	+2	+3	
		A,I,LI, A/B P/C,Ponct, P LN,					
	P						SP
		KZ,JLN,U,B IS/KS LN	Vsa/SP	P V,P/P.V			
	¬ P						Vsa
							Vsa/SP
		'.' B,U PE/CE/BE JKP,JKS, (ADV4) LN		I			
	2X/V YM,U,JLN,PC 3X/V ¬ P						Voa
			V.a/SS	Pa,P.N,P/ P.N			
	A,LZ,LI P Ponct	I,IS,IG, (ADV4) PD/CD/B,P LN P/B					SS
							Voa/SS
		IS					SS
	Voa	',',U.PC 2x/IV E	Vra/Vrp/SS				Vra
							Vrp

Résolution des ambiguïtés : verbales

PREDECESSEURS			AMBIGUITE	SUCCESSEURS			RESULTAT
-3	-2	-1		+1	+2	+3	
		P-V, P/P-V KZ, JKS, (B) (Adv2)					
	JLN, U						Vsa
		Is, (Adv3), (U)		Vsa, X/Vsa			
	Is, (Adv3), (U)	A	Vsa/SS/SP				SS
		Ip, (Adv3), (U)		Vpa, X/Vpa			
	Ip, (Adv3), (U)	A					SP
		I					SS/SP
							Vsa/SS/SP
				PA, PAN, PAA			
	P, J, LZ, P/B PE/BE/CE, X/V	Adv3 Adv3					
	P	'which', (U)	V-a/A-				A-
	JKp, JKs, Y, JLn 3x, 3x/V	(Adv3), (U) (Adv3) (U)					V-a
							V-a/A-

Résolution des ambiguïtés : verbales

DE	PREDECESSEURS -3	PREDECESSEURS -2	PREDECESSEURS -1	AMBIGUITE	SUCCESSEURS +1	SUCCESSEURS +2	SUCCESSEURS +3	RESULTAT
	-	-	I,A,A1		-	-	-	SS
					I,A,A1,P	-	-	VØA
)				YØ,UØA/SS	U	XØ,VØA,XVV		
					U	B	XØ,VØA,XVV	YØ
					-	-	-	VØA/SS
			YØ					
)		YØ	U,B	VØA/2G/P				VØA
								P
			I	VØA/A1/SS	SS,SP A1 VSA/SP			
					E	A1		A1
					VS,YM			
					XS/VS,PA			SS
								VØA
			(B)JKP,JKS					
			(B)PC,YM					VØA
			I,IS/KS		'.',':IT', P,I			
			E	VØA/A/SS				
			A	VØA/A2/SS	VDA/VDP			SS
			I,IS/KS		-			
			SP		P,V			A
			U,C,1X, 1X/V		-	-	-	
				4xD/VDA/VDP/A	P,P/C,P/C/B			
					V-a,Y,X,XV B,'.',',', ')'			VDP

Résolution des ambiguïtés : verbales

E	PREDECESSEURS -3	-2	-1	AMBIGUITE	SUCCESSEURS +1	+2	+3	RESULTAT
			2X,2X/V JK,mS,pS	4xd/Vda/Vdp/A				Vda
			I					A
					'TØ'	VØA,X/V		4Xd
			E					Vda/A
	-	-	-		-	-	-	Vda/Vdp/A
			'TØ'	Voa/A1/B1				VØa
					I			B1
								A1
			1X/1V	'REGARDING'				Vga
					-	-	-	P
	-	-	Ponct	'INCLUDING'				P
	-	-	-		-	-	-	Vga
			','	'DEPENDING'	'ON'			D-p
	-	-	-		-	-	-	VGA
			'('	'PROVIDING'	I			C
								VGA
			'.'	'ASSUMING'				C
								VGA
			1X,1X/V	VGA/A1	I,I/K,P,C P/C,F,'.' C/B			
		¬ A	S,C,C/M, ',' E		KZ	P,'.'		VGA
		I	I,I/K,A1 B					

Résolution des ambiguïtés : verbales

DE	PREDECESSEURS -3	-2	-1	AMBIGUITE	SUCCESSEURS +1	+2	+3	RESULTAT
			P,A		S			
0		A	E	VGA/Al				
					SS			
		.						Al
			1X,1X/V, 'ONE'					
					I,I/K,gK			
					'what', 'that'			
			C		',',',','.',P S,K			
			PF/CF		E1			
					E2	P	SS	
			B,A/B					
			V		¬P, ¬SP			
			';',',','.', ','		¬P, ¬SS			
			P,PF/CF		A	SS		
			P,PF/CF	Vga/SS	S			
		','	E					
			E		S			VGa
			I,I/K,A					
			V/A,V/V/A					
					KZ	¬P, ¬'.'		
					Y,X,V,X/V, LN			
					E1	¬ 'FØR'		
					E2	SS		
					E2	P	SS	
			V		P,SP			
			';',',','.'		P			
			','		SS			SS

Résolution des ambiguïtés : verbales

PREDECESSEURS			AMBIGUITE	SUCCESSEURS			RESULTAT
-3	-2	-1		+1	+2	+3	
		P,PF/CF		S			
		P,PF/CF	VGa/SS	A,B	SS		
	¬ ','	E		S			SS
							VGA/SS
		S,P		I			
		','		'IF'			VGa
		S,P	Vga/A1/SS/P	SP			
		'NEXT'					A1
		I		PONCT,X,V			SS
							P
		U Pa,A, 'ONE'					
		¬ 'FØR'		E1			
				',', Y			SS
		Vga	Vga/A1/SS				
				SP			
		¬ 'ONE'		S			A1
		A1		S			A1/SS
							VGA
				(U),(B), VOA			
			3X/3V	VTA,VRA,XV			3X-
							3VOA
				(U),(B), 'TØ'			
			2X/2V	(CODE1)			2X
							2V-a
				(B) (CODE2),(U)			1XG
			1XG/Vga				1VGA

Résolution des ambiguïtés : verbales

PREDECESSEURS			AMBIGUITE	SUCCESSEURS			RESULTAT
-3	-2	-1		+1	+2	+3	
				(U),(B)	CODE2		
			1X/V	VG ,XG/VG			1X
							1V
	2X/V,2X	(Adv1)(U)	Vna/Vnp				Vna
							Vnp
	2X/V,2X	(ADV1),(U)					
	JK,JLN,LZ	(ADV1),(U)					
Ponct	KZ	(ADV1),(U)					
P, P/C	LN	(ADV1),(U)	V.a/Vp				V.a
	1X,1X/V	(ADV1),(U)					
P,P/C	LN			P			V.p
							Va/Vp
		1X/V,(ADV4)		I			V p
		I,'.', I/K	V.p/A1				
	P,P/C	Ip/Kp/D1 LN					A1
							V.p/A1
		2X,2X/V					
		JK,JL,LZ	Vda/A1	I			Vda
							A1
		2X,2X/V					
		JK,JL,LZ					Vda
		'.'	Vda/A				A
							A1

Résolution des ambiguïtés : verbales

PREDECESSEURS			AMBIGUITE	SUCCESSEURS			RESULTAT
-3	-2	-1		+1	+2	+3	
			Vda/A2	'.'			A2
							Vda
	JK,JLN,LZ	(ADV1),(U)	V a/V p/A1				
	2X,2X/V	(ADV1),(U)					
	¬P, ¬P/C	LN					V.a
	I/K/B,I/'K	(ADV1),(U)		SS			
	I,P,':','.'	(ADV1),(U)					
	P,P/C	LN					A1
		1X,1X/V		P,PF/CF			V p
		'much', 'more'					V p/a1
							Va/Vp/A1
		'.'	Vna/Vnp/A1/P	I			P
		'.'		¬I			A1
							Va/Vp/A1
		1X,1X/1V	Vda/Vdp/A1/SS/B				
	I, A	P,E,PF/CF					Vdp
		Ln					Vda
		¬I		ponct			B
		I		ponct			
	I,A	P,E,PF/CF		V			SS
				S.VoA/SS			A1
							A1/SS

Remarques :

Il ne serait pas très compliqué d'écrire un programme susceptible de saisir ces tableaux sur des cartes de données et de générer les traitements correspondants. Cependant, il reste avant d'en arriver là, quelques problèmes à résoudre, notamment en ce qui concerne le format des données.

Puisque nous avons un code à 6 positions et que l'on peut être amené à tester chacune d'elles, il faudrait étendre chaque code sur 6 caractères. Si une ou plusieurs des positions sont inutiles, il faudrait mettre au point un système permettant de les ignorer.

(123456
Exemple : un verbe peut se coder ␣␣Vsa␣ ou ␣␣Vda␣ ... Si l'on veut tester l'hypothèse verbe, il faudrait indiquer de l'une des manières suivantes que c'est le caractère 'V' en position 3 qui est à tester :

1) ... V ... (indiquant que les p.1,2,4,5,6 sont à ignorer)

2) V(p3) : tester 'V' dans la position 3.

Dans les deux cas, il en résulterait une grande perte de place. Faute de mieux, dans l'immédiat, nous nous contentons de ces tableaux qui permettent de programmer les traitements assez vite.

III.1.6. Représentation formelle des règles de levée d'ambiguïté.

III.1.6.1. Introduction :

On a vu comment il était possible de présenter sous un format agréable des tableaux permettant d'indiquer de quelle façon les diverses ambiguïtés vont se lever.

Afin d'utiliser ce modèle, il nous faut procéder à une réorganisation du dictionnaire.

Il est possible de remplacer le code existant par un code à 3 caractères alphanumériques. Le premier représentera la nature grammaticale du mot, les deux suivants un numéro d'ordre dans la liste des différents types de mots rencontrés.

Ainsi les verbes vont recevoir les codes :

V01, V02 ... en commençant par les verbes non ambigus.

Le code ainsi constitué est bien entendu moins riche que celui qu'il remplace ; en effet on n'y verra pas figurer les catégories grammati-

cales, ni les rections mais l'essentiel est qu'il soit suffisant pour la levée d'ambiguïté. Un programme de transcodage peut facilement faire la conversion.

III.1.6.2. Transformation du code.

Le tableau suivant représente le nouveau code en fonction du code ancien. Les conventions d'écriture suivantes sont employées :

- le ";" indique une fin de code
- le "/" indique une ambiguïté
- une "," indique un choix entre plusieurs codes anciens pour un code nouveau unique.

Codes nouveaux	CODES ANCIENS
A01	A ;
A02	A1 ;
A03	A2 ;
A04	A / SS ;
A05	A1 / SS ;
A06	A2 / SS ;
A07	A / B ;
A08	A / SS / B ;
B01	B
B02	BZ
B03	HB
C01	C ;
C02	TC /
C03	TC / M ;
C04	TCW / MW ;
C05	CH / B ;
C06	C- ;
C07	CT / IS / KS / LT ;
D01	D2
E01	E
F01	F
I01	I , IG
I02	IS ;
I03	IP ;

Codes nouveaux	Codes anciens
I04	I / B ;
I05	-I- / -K- / -B- ;
I06	I- / K- / A- / B ;
I07	I- / K- / D- / B ;
I08	I / K / SS / B ;
I09	I- / K- ;
I10	I- / K- / D- ;
I11	I- / SS / B ;
K01	KS , KP
K02	KZ
K03	JKS , JKP
L01	LN , LK , LJ ;
L02	LT
L03	JLN ;
M01	MH ;
M02	MW ;
P01	U PC ;
P02	- P- ;
P03	P- / B ;
P04	U PD / U TCD / 2B ;
P05	P- C-
P06	P / A1 / SS / B ;
P07	P / A1 / B ;
P08	P C B
S01	SS ;
S02	SP ;
S03	PSSA ;
S04	MSS ;
U01	U
V01	VDA ;
V02	VDP ;
V03	VGA ;
V04	VMA ;
V05	VNA ;
V07	VNP ;
V08	VOA ;
V09	VPA ;
V10	VRA ;
V11	VRP ;
V12	VSA ;
V13	VSP ;

Codes nouveaux	Codes anciens
V20	VSA / SP ;
V21	V-A / SS
V22	VRA / VRP / A / SS ;
V23	VRA / URP / SS ;
V24	VSA / SS / SP ;
V25	V-A / A ;
V26	VOA / P
V27	VOA / A / SS ;
V28	VOA / A / B ;
V29	VGA / P ;
V30	VGA / C ;
V31	VGA / A1 ;
V32	VGA / SS ;
V33	VGA / A1 / SS / P ;
V34	VGA / A1 / SS ;
V35	V-A / V-P ;
V36	VDA / VDP / C ;
V37	V-P / A1 ;
V38	VDA / A- .
V39	VDP / SS
V40	V-A / V-P / A1 ;
X01	1XD / 1VDA ;
X02	1XG / 1VGA ;
X03	2XG / 2VGA ;
X04	3XM / 3VMA ;
X05	1XO / 1VOA ;
X06	2XO / 2VOA ;
X07	3XO / 3VOA ;
X08	1XP / 1VPA ;
X09	1XS / 1VSA ;
X10	2XS / 2VSA ;
X11	4XD / 4VDA / 4VDP / A1 ;
Y01	Y , YM , YO ;
Y02	YO / VOA / SS

RECAPITULATIF

TYPE	AMBIGUS	NON AMBIGUS	TOTAL
A	5	3	8
B	0	3	3
C	4	3	7
D	0	1	1
E	0	1	1
F	0	1	1
I	8	3	11
K	0	3	3
L	0	3	3
M	0	2	2
P	6	2	8
S	0	4	4
U	0	1	1
V	28	13	41
X	10	0	10
Y	1	1	2
TOTAUX	62	44	106

III.1.6.3. Réalisation

III.1.6.3.1. Format des données.

Il nous faut en premier lieu définir sous quelle forme les chercheurs linguistes pourront présenter les règles de levée d'ambiguïté. Après avoir établi les tableaux de levée sous le format défini plus haut, ils auront à choisir dans la liste des nouveaux codes leurs symboles, ils auront à définir l'ambiguïté à lever en prenant son code dans cette même liste ; une règle est un ensemble de symboles terminée par un point-virgule. Nous ferons une règle par résultat possible et par type de contexte. Il faudra donc indiquer par deux chiffres (longueur du contexte gauche, longueur du contexte droit) l'importance de ce contexte.

Enfin une règle aura le format suivant :

```
    SYM AMAV / id1 / id2 / ... / idn/, /id'1/ id'2 ... *Sym*...
1           * RES ;
```

SYM : symbole de l'ambiguïté à lever
AM : nombre mot 'avant' utilisés par la règle (en amont)
AV : nombre mot 'après' utilisés par la règle (en aval)
id1... idn : un choix de code pour les prédécesseurs ou successeurs
RES : le résultat choisi si cette règle est validée
1 marque d'une carte suite

Exemple : Le fait que l'ambiguïté V21 (V.a/SS) située en tête de phrase (précédée de '.') et suivie d'un artida (I01) se résoud par V a (V08) sera symbolisée par la règle :

V21 1 1 . * V21 * I01 * V08 ;

III.1.6.3.2. Méthode de traitement.

Le processus que nous mettons en oeuvre va donc comporter deux phases bien distinctes :

- une phase de compilation des règles qui consistera à les ranger en mémoire sous une forme pratique, à contrôler leur syntaxe.
- une phase d'exploitation de ces règles en vue de la levée d'ambiguïté.

La transmission des données d'un programme à un autre se fera par un fichier à enregistrements de longueur variable ; chaque enregis-

trement contiendra donc la longueur d'une règle suivie de l'image de son implantation en mémoire.

III.1.6.3.3. Rangement des règles en mémoire :

Pour ranger les règles en mémoire, nous proposons le format suivant :

4 mots : H, H + 1, H + 2, H + 3 contenant respectivement aux adresses

H : le sigle de la règle
H + 1 : son encombrement (nombre mot)
H + 2 : son résultat
H + 3 : 1er 1/2 mot le nombre AM (amont)
2e 1/2 mot le nombre AV (aval)

L = AM + AV définit la longueur du contexte à explorer.

Pour chaque possibilité de la règle don aura donc un groupe de quatre mots contenant successivement les codes des différents mots du contexte.

L'encombrement en mots-mémoire d'une telle règle est 4 + k L où k est le nombre de choix possibles dans les contextes ; (4 mots suivis des KL mots définissant la règle).

Exemple :

Considérons la règle V20

V20 20 $\underbrace{\text{P01/P02}}_{2}$, $\underbrace{\text{A01/A02/A03/I01/I02/I03 .../PON*V20*S02}}_{12}$;

La valeur de k est donc 24 (2 x 12) et L vaut 2. La règle V20 occupera donc 52 mots en mémoire répartis ainsi :

H H+4

V20	0024	S 02	2 0	P01	A01	P01	A02
P01	A03	P 01	I 01	P01	I02	P01	I03
						H+48	H+49
						P02	PON

Nous allons de plus créer en mémoire deux tables récapitulatives.

a) La table récapitulative des règles (REG-RECAP)

La ligne n° i de la table occupe 5 mots.

sigle	encombr.	adresse	AM	AV

(pour la règle n°i)

Sigle : l'ambiguïté que cette règle peut lever

Encombrement : le nombre de mots-mémoire qu'elle occupe

Adresse : l'adresse d'implantation de cette règle

AM, AV : les dimensions du contexte à explorer

Exemple : si la règle V20 est la règle n° 30

on trouvera à la ligne 30 :

V20	52	S02	2	0

b) Table récapitulative des ambiguïtés (AMB-RECAP)

chaque 'ligne' occupera deux mots

Sigle : le code de cette ambiguïté

n1 : les règles concernant cette ambiguïté sont récapitulées dans la table des règles à partir de sa n1 ligne

n2 : il y a n2 lignes dans cette dernière table qui décrivent les règles levant cette ambiguïté.

Exemple : si l'ambiguïté V20 est la 3e dans notre liste et peut se lever par 10 règles distinctes (n° 30 à 39) la table AMB-RECAP contiendra en ligne 3

V20	30	10

III.1.6.3.4. Exploitation des règles sur un texte :

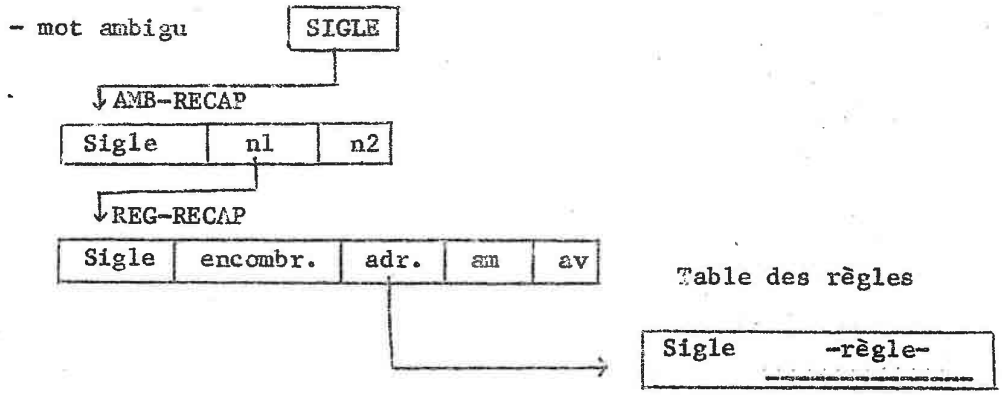
On suppose les règles ramenées en mémoire et on se trouve en présence d'un texte codé par le nouveau code à trois positions.

La rencontre d'un élément Vnn provoque une recherche dans la table AMB-RECAP qui va nous indiquer que les règles de levée correspondantes se trouvent récapitulées dans la table REG-RECAP entre les lignes n1 et n1 + n2 - 1.

On parcourt les lignes de la table. On sait aussitôt s'il est pos-

sible de tester le contexte (compte-tenu du rang du mot dans la phrase).

On emploie donc un chainage à trois niveaux :



Les quantités AM et AV étant connues et valides (i.e. la possibilité de créer le contexte existe) on peut créer le contexte sous la forme : cod (mot AM) I cod mot (am-1) ... ' code mot (AV)
et c'est cette chaine de caractère que l'on va aller comparer aux divers éléments du modèle.

III.1.6.3.5. Résultat :

Si ce test s'avère positif, le nouveau code à affecter au mot ambigu est indiqué au début de la règle considérée. Sinon il nous faut tester successivement toutes les règles concernant ce sigle jusqu'à épuisement de la liste ou jusqu'à un résultat positif. Si aucun modèle ne s'est révélé conforme au texte, le mot demeure ambigu et son code n'est donc pas modifié.

III.1.7. Vers le groupe verbal.

III.1.7.1. Pronoms et adverbes ambigus.

Les ambiguïtés verbales étant, dans leur majeure partie levées, nous sommes en mesure de reconnaître l'environnement verbal non ambigu de certains adverbes ou pronoms. Deux règles de grammaire peuvent être ainsi énoncées.

1°/ Impossibilité à tout autre terme qu'à un pronom d'être inséré entre un verbe et son adverbe.

2°/ Certains types d'adverbes peuvent s'insérer dans des groupes verbaux.

En conséquence, si une ambiguïté porte dans ses codes la possibilité pronom, si elle est située entre un mot codé x, y ou v (auxiliaire, défectif ou verbe) non ambigu et un adverbe appartenant à une liste bien définie, elle sera levée au profit du seul code pronom. Si un adverbe ambigu appartient à cette même liste et s'il est situé entre un mot codé x ou y (auxiliaire ou défectif) et un verbe (codé v), il deviendra adverbe non ambigu.

III.1.7.2. Conclusion :

Le verbe et les auxiliaires sont à présent suffisamment bien définis pour nous permettre de construire le groupe verbal autour d'eux. Là encore nous utiliserons des règles de succession ou à l'inverse d'impossibilité de contexte pour en définir les frontières.

III.1.8. Résultats chiffrés :

Nous avons testé l'efficacité des programmes de levée d'ambiguïtés en les regroupant par degré : ambiguïtés à 2, 3 ou 4 et 5 termes et cela sur deux textes.

1°/ Les "33 phrases"(échantillon tiré du corpus entier).
2°/ Le livre 13: le 13e article de notre corpus.

33 phrases

d° d' ambiguïté	occurrences	levées totales	levées partielles	non levées
2	124	110	0	14
3	20	16	1	3
4 et 5	6	5	0	1

Verbes non ambigus : 40
nombre total d'éléments présentant au moins 1 analyse verbe : 190
nombre total de mots : 1100
nombre total d'analyses :1500

Livre 13

d° d' ambiguïté	occurrences	levées totales	levées partielles	non levées
2	135	117	0	18
3	28	23	2	3
4 et 5	5	5	0	0

Nombre de verbes non-ambigus 43

Nombre total d'éléments présentant au moins une analyse verbe : 211

Nombre total de mots : 1600

Nombre total d'analyses : 2100.

III.2. RECONNAISSANCE DES GROUPES VERBAUX.

III.2.1. Définition

Nous entendons par groupe verbal toute suite de mots introduite par un verbe ou un auxiliaire, et jouant, dans la proposition où il se trouve, le rôle de verbe et seulement celui-là.

Le groupe verbal a toujours un ou plusieurs sujets, il peut avoir des objets, des attributs ou des circonstants. Il est à noter que le sens que nous donnons ici au terme de groupe verbal est un sens beaucoup plus restrictif que celui que les linguistes attribuent à la notion de syntagme verbal.

Pour le linguiste en effet le syntagme verbal est ce qui, dans la phrase est porteur de sens et par conséquent le prédicat (objet, attribut) est inclus dans le groupe.

III.2.2. Ce que l'on veut obtenir.

La construction du groupe verbal étant supposée réalisée, étudions quels renseignements nous seront nécessaires pour poursuivre l'analyse. En d'autres termes, quels codes allons-nous attribuer aux groupes verbaux construits et quels liens lient ces codes aux constituants des groupes ?

Le groupe verbal sera muni en premier lieu d'un code grammatical et en second lieu d'un code syntaxique.

III.2.2.1. Code principal (grammatical).

C'est un code dont les six positions (P1, ... P6) peuvent prendre les valeurs suivantes :

P1 : U le groupe comporte un seul mot
D le groupe comporte plusieurs mots

P2 : 'b', 1, 2, 3 la plupart du temps cette position est blanche sauf si le groupe est introduit par des auxiliaires bien particuliers
1 : auxiliaire BE
2 : auxiliaire HAVE
3 : auxiliaire DO

P3 : V la valeur V indique la nature verbale du groupe

P4 : I le nombre est indéterminé
P le nombre est pluriel
S le nombre est singulier
G pas de nombre ; c'est un groupe en ING

PS : C forme conjuguée
N forme non conjuguée

P6 : A le verbe est employé à la voix active
P le verbe est employé à la voix passive.

On peut observer que ce sont là tous les renseignements qui peuvent être découverts par le lecteur au seul examen du groupe, et sans faire appel au contexte. Il n'en est pas de même pour les codes syntaxiques.

III.2.2.2. Codes syntaxiques :

Nous mettons là les catégories grammaticales, les rections et les insertions.

Les catégories grammaticales et les rections affectées au groupe verbal reconnu sont celles de son dernier terme.

Les insertions sont les codes des termes non verbaux insérés dans le GV ; il s'agit essentiellement d'adverbes ou parfois de pronoms.

III.2.3. Formes simples et formes composées :

Tous les mots issus de la levée d'ambiguïté et présentant l'un des codes : VØ ou VR ou VT acquièrent le code ambigu ⊔ VICA / ⊔ VINA ; en effet, ces trois codes représentent aussi la marque de l'infinitif, d'où l'ambiguïté entre forme conjuguée et forme non conjuguée.

Tous les verbes codés Vg deviennent ⊔ Vg-a

Une forme simple comporte donc un seul mot graphique et est issue d'un verbe codé précédemment d'une des quatre façons signalées plus haut.

III.2.4. Relevé des formes composées.

III.2.4.1. Formes à deux termes (groupes verbaux à deux mots)

. X - V (auxiliaire suivi de forme verbale).

Les deux termes sont regroupés et le code évolue comme suit :

2XO Vna ⟶ D Vica / D vina (auxiliaire have)
1Xg Vdp ⟶ D Vgnp
1Xs Vdp ⟶ D Vscp (auxiliaire BE)
1Xp Vdp ⟶ D Vpcp

3XO Voa
3Xm Voa ⟶ } D Vica (auxiliaire DØ)

. Y - V (défectif suivi de forme verbale)

Ym . Voa ⟶ D Vica (toutes les formes du futur)

III.2.4.2. Formes à trois termes verbaux :

. Y - X - V (auxiliaire au futur ou conditionnel suivi d'une forme verbale)

Quels que soient les auxiliaires en présence, le code résultat sera D Vicp

. X - X - V (auxiliaire à un temps composé suivi d'une forme passive).

2Xs 1Xn Vdp ⟶ D Vscp
2Xd 1Xn Vdp ⟶ D Vicp
2Xo 1Xn Vga ⟶ D Vica / D Vina
2Xo 1Xn Vdp ⟶ D Vica / D Vina
4Xd 1Xo Vdp ⟶ D Vicp

III.2.4.3. Autres formes de groupes verbaux.

Toutes les formes citées dans III.2.4., c'est-à-dire non réduites à un mot, peuvent admettre, un ou plusieurs termes non verbaux insérés.

Ainsi verra-t-on des chaines du type :

X B V , Y B V , X X B V etc...

III.2.5. Exemples :

	MOT 1	MOT 2	MOT 3	MOT 4	MOT 5	GROUPE
termes	ARE	DISGUSSED				are discussed
codes	1Xp	Vdp				D VPCP
termes	DID	MATURE				did mature
codes	3Xm	Voa				D VICA
termes	MUST	BE	LEFT			mist de Left
codes	Ym	1Xo	Vdp			D VICP
termes	HAS	BEEN	GREATLY	RESTRICTED		
codes	2Xs	1Xn	B (adverbe)	Vdp		D VSCP (B)
termes	SHOULD	HAVE	BEEN	SOMEHOW	RENOVED	Should...removed
codes	Ym	Voa	1Xn	B	Vdp	D VICP

III.3. CONSTRUCTION DES GROUPES VERBAUX.

III.3.1. Introduction

La construction des groupes verbaux relève d'une démarche extrêmement simple. Nous disposons en mémoire de la suite des mots de la phrase avec leurs codes et particulièrement de la liste des verbes (devenus à présent non ambigus).

Les règles de formation du groupe verbal, mises en évidence par Mademoiselle Lecomte dans sa thèse de 3ème cycle (C.R.A.L. n° 17) nous permettent de concaténer des suites de mots en groupes verbaux.

On peut citer en exemple des règles telles que :

- auxiliaire suivi de auxilié ⟶ groupe verbal
- verbe conjugué seul ⟶ groupe verbal.

La nature des composants de ces groupes permet alors de leur affecter les codes convenables. C'est l'algorithme qui découle naturellement du processus de levée d'ambiguïtés.

Sur les textes ce procédé nous a donné des résultats appréciables en ce sens que nous n'avons obtenu ni résultat erroné ni groupe non formé. Cependant, ce n'est pas là une méthode pleinement satisfaisante, tant par la lourdeur de sa mise en oeuvre (nécessite un gros travail préparatoire sur les ambiguïtés) que par la difficulté que nous avons à la formaliser.

C'est pourquoi nous allons nous attacher à mettre en lumière un procédé qui évite les levées d'ambiguïtés et qui construit les groupes verbaux "à priori".

III.3.2. Représentation dans une C. Grammaire

III.3.2.1. Exemple :

Le groupe verbal HAD BEEN SOMEHOW REMOVED reçoit pour chacun de ses éléments dans le dictionnaire les codages suivants :

HAD	2Xd, Vda	(ambigu auxiliaire, verbe)
BEEN	1Xn, Vna	(ambigu auxiliaire , verbe)
SOMEHOW	B	(adverbe)
REMOVED	Vda, Vdp	(ambigu actif, passif)

On perçoit bien que cette forme n'a rien d'ambigu. Pourtant trois de ses termes le sont. Pourquoi ?

HAD est un verbe actif dans le contexte : he had two boys
mais un auxilaire dans : he had eaten
de même dans This was removed REMOVED est un participe passé, tandis qu'il est un verbe conjugué actif dans : the man removed the piece.

La présence de contextes significatifs va permettre à nos algorithmes successivement :

- de lever les ambiguïtés :

Had	BEEN	SOMEHOW	REMOVED
2Xd	1Xn	B	Vdp

- de reconnaître que la chaîne

2Xd - 1Xn-B-Vdp autorise un regroupement verbal avec l'attribution du code : D VICP.

III.3.2.2. Le but :

L'idée consiste à éviter l'étape de levée d'ambiguïté et à construire directement le groupe à partir des termes ambigus.

Pour ce faire, nous allons formaliser dans une C. Grammaire les règles de constitution du groupe verbal et tenter ensuite de faire une analyse syntaxique sur certains termes des phrases.

III.3.2.3. La syntaxe du groupe verbal.

(GV axiome)

GV :: = GVU GVD (groupe à un ou plusieurs termes)

GVU :: = VIC VGN VIN VCA

VIC :: = VIC1 VIC2 VIC3 VIC4 VIC5

VIC1 :: = ppf1 Axb1

VIC2 :: = Axa1 Vsi1

VIC3 :: = ppf4 pph1

VIC4 :: = Dcc2 vsi2

VIC5 :: = Vsi3 ppf3

VGN :: = Pgg1 Pgg2 Pgg3 Pgg4 Pgg5 Pgg6

Vin :: = ppf2

VCN :: = Axa2 Axa3 Vsj1 Vsj2 Vsj3

ici se terminent les règles du groupe verbal à un terme ; ce qui jusqu'à présent s'exprimait pour nous plus simplement par : mot dont le code principal présente en position 3 l'une des lettres : X, Y, V :

(1) GVD :: = GVi GVC

(2) GVi :: = GViC1 GVIC2 GVIC3 GVIC4 ;

(3) GVC :: = GVC1 GVC2 GVC3 GVC4 ;

(4) GVIC1 :: = De . Ins. VSi Axc2 . ins . Vsi

(5) De :: = Dcc1 Dcc2

(6) Vsi :: = Vsi1 Vsi2 Vsi3

(7) GVIC2 :: = axc2 . ins . ppfh

(8) GVIC3 :: = DDE . ins . Axa1 . ins . ppfh

(9) DDE :: = Ded Axd1

(10) PPfh :: = PPf PPh

(11) PPf :: = PPf1 ..... PPf6

(12) PPh :: = PPh1 PPh2 PPh3

(13) GVIC4 :: = Axb2 . ins . Axa1 . ins . PPfh

(14) GVC2 :: = Axa2 . ins . PPfh

(15) GVC3 :: = AxD3 . ms . Pta / Axa3 . ins . Pgg
(16) Pta :: = PPfh PGg
(17) PGg :: = PGg1 ..... PGg6
(18) GVC4 :: = AxD1 . ins . Axa1 . ins . PPfh
AxD3 . ins . Axa1 . ins . PPfh
(19) ins :: = B S K

tous les symboles terminaux, c'est-à-dire ne figurant à gauche dans aucune des règles sont directement accessibles dans le dictionnaire.

Cette méthode résoudra parfaitement le problème de l'ambiguïté des termes dans les groupes verbaux à plusieurs mots. Par contre, si l'on voulait également lever les ambiguïtés sur les termes donnant naissance à un GV1, il faudrait inclure celui-ci dans un ensemble syntaxique plus large.

III.2.4. Résultats "dans les textes"

Nous donnons ici, à titre indicatif, le nombre de chaines verbales, classées par nombre de mots graphiques reconnus ou restant ambigus et faisant partie de notre corpus-test.

Texte	Nombre de mots	Chaines d'un mot		chaines de 2 mots		chaines de 3 mots et plus		Total chaines
		reconnues	ambigues	reconnues	ambigues	reconnues	ambigues	
33 phrases	1260	38	16	17	2	21	2	96
Livre 13	1540	61	33	31	3	22	0	150
Livre 01	2580	115	45	46	9	28	3	246
Total		214	94	94	14	71	5	492

Chapitre IV

LA CHAINE NOMINALE

IV.1. AMBIGUITES EN SUSPENS

La logique veut qu'après le groupe verbal nous nous attachions au groupe nominal. Arrêtons-nous quelques instants, après ces trois premiers chapitres, aux résultats du traitement qui, des codages préliminaires, nous a conduit aux groupes verbaux. Nous disposons à présent d'un texte qui ne comporte presque plus d'ambiguïtés verbales. Nous remarquons que les ambiguïtés qui demeurent concernent, pour la plupart, trois classes de mots :

les introducteurs, les adjectifs, les prépositions.

Ces trois classes sont particulièrement proches du groupe nominal. En effet si l'on appelle introducteur tout ce qui est article, déterminant, on a trouvé là la frontière gauche du groupe nominal, si l'on considère les adjectifs, ce sont les centres de la plupart des groupes nominaux : le schéma introducteur - adjectif - substantif étant très fréquent. Enfin, dans une phrase les prépositions sont les signes annonciateurs des groupes nominaux.

Comme on le verra plus loin, mais ce n'est que pure convention, pour nous les prépositions ne font pas partie du groupe nominal.

Outre ces trois grandes classes, on trouve en marge quelques ambiguïtés portant sur des groupes verbaux a un seul mot, que l'on n'a pas su lever auparavant.

- liste des ambiguïtés à lever avant le regroupement nominal.

IV.1.1. Classe des introducteurs :

Nous indiquons ici les codes des mots ambigus appartenant à la classe des introducteurs :

IP / KP / DI	introducteur, pronom coordonnant
I / SS	introducteur ou substantif
I / K / A	introducteur pronom ou adjectif
I / K / SS / B	introducteur pronom ou substantif ou adjectif
I / B	introducteur ou adverbe
I / A / B	introducteur, adjectif ou adverbe
I / K / B	introducteur pronom ou adverbe
I / K	introducteur ou pronom

Deux mots, en raison de leur fonctionnement vraiment particulier ont dû recevoir un traitement spécifique. Ce sont :

LITTLE codé I / K / A ou B

et THAT codé CT / IS / KS / LT

L'ambiguïté la plus fréquemment rencontrée est sans conteste l'ambiguïté I/K, c'est-à-dire introducteur ou pronom.

IV.1.2. Classe desadjectifs :

Ici nous trouvons trois ambiguïtés seulement ; ce sont :

A / B	adjectif ou adverbe
A / SS	adjectif ou substantif
A / SS / B	adjectif ou substantif ou adverbe.

IV.1.3. Classe préposition :

P / A / SS / B	préposition, adjectif substantif ou adverbe
P / B	préposition ou adverbe
P / C	préposition ou conjonction
P / C / B	préposition ou conjonction ou adverbe.

IV.1.4. En marge :

En marge de ces trois grandes classes nous trouverons les résidus des traitements précédents portant sur les verbes.

Vc / V.n	Verbe conjugué ou non conjugué
Vg / SS	Verbe en ing ou substantif
V / V / A	verbe, verbe ou adjectif
V / SP	Verbe ou substantif.

IV.2. LEVEE DES AMBIGUITES PREPARATOIRE A L'ETABLISSEMENT DES GROUPES NOMINAUX.

La levée des trois grandes classes d'ambiguïtés d'une part et des ambiguïtés que nous avons appelées "en marge" d'autre part, ne diffère pas ou peu des méthodes employées lors du travail préparatoire à la constitution des groupes verbaux. C'est pourquoi nous ne nous attarderons pas sur la démarche suivie.

Par contre, nous présentons les tableaux récapitulatifs d'un certai nombre de ces différentes règles. Ces tableaux appellent au moins une remarque. Il s'agit de leur simplicité, plus grande que lors de la levée des ambiguïtés du verbe.

Résolution des ambiguïtés : classe des prépositions et conjonctions

DE	PREDECESSEURS -3	-2	-1	AMBIGUITE	SUCCESSEURS +1	+2	+3	RESULTAT
			I	P/A1/SS/B	S			A1
			I		¬ S			SS
					Ponct,W			B
								P
			I	P/A1/B	S			A1
					V i NP			B
								P/B
			P	PP/B,Pr/B	F,I,Q,'.' 'one-quarter'			B
								P
			P-		Ponct,P P/C,P/C/B V-C,W			
					E	V,¬P/B		B
								P
			'either'	PD/CD/2B	V g			
					A,B			PD
					P,',',W,C (B),JK,V-c Ving			CD
			P		A,B			
					A,B	'AS'		2B
		2I,4I	S					PD/CD
								PD/CD/2B
					Ponct,C			
					E	¬P		B

Résolution des ambiguïtés : prépositions et conjonctions

CODE	PREDECESSEURS -3	-2	-1	AMBIGUITE	SUCCESSEURS +1	+2	+3	RESULTAT
360				P/C/B	E	P		B
								P/C
			','	PF/CF				PF
					P,JK,W,',' C			P/C
								C
370					(B), V-g, GK, Ln			
			'either'	P-/C-	E,F			
		P	E					P
								P/C
90				C/B	JK, KS			C
								B
80				CH/B	A,B			CH
								B

Résolution des ambiguïtés : verbes et groupes verbaux (en marge)

CODE	PREDECESSEURS -3	PREDECESSEURS -2	PREDECESSEURS -1	AMBIGUITE	SUCCESSEURS +1	SUCCESSEURS +2	SUCCESSEURS +3	RESULTAT
240	¬P, ¬P/C	Ln	JK-, Adv4, U JLN, LZ 'more', 'much',	-Vic/Vin	'that', ¬P			-Vic
			U- PC					-Vin
								-Vic/-Vin
245	¬P, ¬P/C	Ln	JK, Adv4, U JLn, LZ 'more', 'much'	U-Vica/U-Vina/ U-Vinp				
	Ln	'Let'	KZ					Vic
			P					-Vinp/-Vina
								U-Vica/U-Vina/U-Vinp
255			Ig	U-Vgna/SS	V-C Ponct			SS
			Ig		Ponct			Sg
								U-Vgna /SS
260			C	U-Vica/U-Vina/ A				A
			'more', 'much' P, PF/CF					U-Vin/A
								U-Vica/ U-Vina/A
265			P, PF/CF	U-Vica/U-Vina/ SS	V-C			SS
					'ye'			U-Vica
			P, PF/CF					U-Vina/SS
								U-Vica/U-Vina/SS

Résolution des ambiguïtés : groupes verbaux et verbes

CODE	PREDECESSEURS -3	PREDECESSEURS -2	PREDECESSEURS -1	AMBIGUITE	SUCCESSEURS +1	SUCCESSEURS +2	SUCCESSEURS +3	RESULTAT
70			'more', 'much'	U-Vsca/sp	V-C			Sp
		P, PF/CF	Ln					U-Vsca

Résolution des ambiguïtés : de la classe des adjectifs

CODE	PREDECESSEURS -3	-2	-1	AMBIGUITE	SUCCESSEURS +1	+2	+3	RESULTAT
10				A1/SS	SS,A,I,K A/SS, V/SS, V/V/SS ',' '.',P,V A2,F ','	A SS		A1 SS A1/SS
10 00			I/K/B,V	A /SS A2/SS	SS,A,I,K P-A			A-
			I,A,P		PA,F,A2			SS
								A-/SS
U			I,A,P	A/SS/B	',' P,'.' E V/V/SS	A/SS/B		SS
			¬ I,¬A, ¬ P		A,SS, A/SS V/SP ','			A B
								A/SS/B

Il est remarquable également qu'à l'examen des résultats, les mots non résolus soient beaucoup moins nombreux que dans la levée précédente.

Cette constation n'étonnera pas. En effet, il est beaucoup plus facile de lever des ambiguïtés lorsque le contexte verbal est bien connu.

Ainsi on peut donner l'exemple suivant :

L'ambiguïté introducteur/pronom (ex. this, there) située devant un groupe verbal non-ambigu se résout immédiatement en introducteur.

Par contre, si avant de lever les ambiguïtés sur le verbe, on avait cherché à lever l'ambiguïté introducteur/pronom dans un contexte où elle était suivie d'un mot ambigu : verbe/substantif, cette résolution se serait avérée impossible.

Les linguistes ont constaté dans cette levée d'ambiguïtés un certain nombre d'erreurs. Ce nombre reste, malgré tout, très limité. Nous avons la certitude que la méthode des élargissements successifs à des textes nouveaux nous permettra de corriger à la fois notre analyse et les programmes pour fournir une levée d'ambiguïté totalement dépourvue d'erreurs.

IV. 4. LES CONSTITUANTS DU GROUPE NOMINAL.

IV.4.1. Définition

On appelle groupe nominal une séquence non disjointe de un ou plusieurs mots contigus susceptibles d'assumer dans la phrase les fonctions suivantes : sujet, objet, circonstant ou complément de nom.

IV.4.2. Limites du groupe

Afin que cette définition représente sans ambiguïté ce que nous entendons par groupe nominal, il nous faut préciser que toute préposition est exclue d'un tel groupe. Par conséquent la structure suivante :

groupe nominal + préposition + groupe nominal
ou groupe nominal + complément du nom

ne constitue pas, pour notre définition, un groupe nominal ; par contre un pronom seul peut constituer un GN.

IV.4.3. Principes de reconnaissance

Comme nous l'avons fait pour le groupe verbal, nous procédons de manière linéaire. Ainsi, la construction se fera par lecture séquentielle des termes de la phrase. Cette façon de procéder est en effet autorisée par notre définition qui interdit au groupe nominal d'être scindé. La rencontre, au cours de la phrase d'un mot susceptible d'introduire un groupe nominal, déclenchera la procédure de construction. De la même manière la rencontre d'un terme de fin de groupe nominal déclenchera l'arrêt de la procédure et l'affectation au groupe nominal du code approprié.

IV.4.4. Code du groupe nominal

Le groupe nominal reçoit lors de sa construction un code à 6 positions qui peut avoir les formes suivantes :

position 1	U	le groupe comporte un seul substantif
	blanc	le groupe ne comporte pas de substantif (pronominal)
	D	le groupe comporte plusieurs substantifs
position 2	1 ou 2	le groupe comporte une comparatif
	blanc	autres cas

position 3 ——→ N marque du groupe nominal

position 4 →S nombre singulier
position 4 →P nombre pluriel
position 4 →blanc nombre indéterminé

positions 5, 6 inutilisées.

IV.4.5. Bornes du groupe nominal

On distingue entre bornes droites et bornes gauches et parmi ces dernières entre bornes incluses dans le groupe nominal et bornes exclues du groupe nominal.

IV.4.5.1. Bornes gauches incluses.

Ce sont les mots dont le code grammatical est I, A, F ou S.

Dans la catégorie I nous trouverons : les introducteurs : articles, déterminants, dans la catégorie F les cardinaux, dans la catégorie A tous les adjectifs sauf ceux qui ne peuvent être qu'attributs, dans la catégorie S tous les substantifs. On a décidé de faire figurer cette dernière catégorie afin de permettre à un substantif quelconque de former à lui tout seul un groupe nominal.

IV.4.5.2. Bornes gauches exclues

Nous avons rangé là toutes les prépositions qui sont annonciatrices du groupe nominal.

IV.4.5.3. Bornes droites

La borne droite est toujours un substantif. Si plusieurs substantifs se succèdent, c'est alors le dernier de la chaine qui forme la borne droite.

IV.4.6. C. Grammaire du groupe nominal

Le groupe nominal (GN) peut avoir un terme (règles 2, 3, 4) ou plusieurs termes (règles 1, 5, 6 ... 11) ce que nous désignerons par GNU et GND comme nous l'avions fait pour le groupe verbal.

```
( 1)  GN    :: = GNU / GND
( 2)  GNU   ::=  S   / K            (substantif ou pronom)
( 3)  S     :: = NS  / SS / SP      (nom propre, singulier, pluriel)
( 4)  K     :: = KI  / KS / KP      (pronom indéfini, singulier, pluriel)
( 5)  GND   :: = GI  . GS
( 6)  GI    :: = I   . GA           (groupe introducteur)
( 7)  I     :: = IO / IS / IP /
( 8)  GA    :: = A  / A. GOR . GA /
( 9)  A     :: = A  / AI            (adjectif quelconque ou spécifiquement
                                     épithète)
(10)  GS    :: = S. GS / S          (partie-substantif)
(11)  COR   :: = 'AND' / 'OR' / ',' /
```

IV.4.7. Exemples de groupes nominaux

Dans le tableau qui suit, nous indiquons l'état du texte avant et après le regroupement nominal. Les groupes constitués apparaissent soulignés.

AVANT		APRES	
Groupes	Codes	Groupes	Codes
This	Ks	This	U NS
may be taken	D NICP	may be taken	D VICP
as	PNV	as	PDV
suggesting	U vgna	suggesting	U Vgna
that	UCT	that	U CT
a	IS	a portion	D NS
portion	SS		
of	PA	of	PA
the	I	the chondroïtin sulfate	D NS
chondroïtin	SS		
sulfate	SS		
had been removed	D VICP	had been removed	D VICP
from	P	from	
the	I	the cartilage	D NS
cartilage	SS		

Le groupe nominal : this est obtenu par application des règles (2) et (4).
Le groupe nominal : a portion est obtenu par application des règles (5), (6), (7), (8), (10) et (3).
Le groupe nominal : the chondroïtin Sulfate est obtenu par application des règles (5), (6), (7), (8),(10), (10), (3).
Le groupe : the cartilage s'obtient de la même façon que le groupe a portion, c'est-à-dire après application des règles (5), (6), (7), (8), (10) et (3).

Comme on le verra dans les exemples qui suivent et ainsi que cela est indiqué par la règle (8) les adjectifs figurant dans le groupe nominal peuvent être coordonnés.

intricate, fine-textured dissection
arid and semi-arid regions
provisional, empirical character

IV.4.8. Conclusion

Dans les groupes constitués à ce niveau on n'a inclu aucun démarcateur de proposition. Il va donc nous être possible de rattacher à chaque proposition un ensemble de groupes nominaux et verbaux de manière unique.

L'idée est de numéroter les groupes nominaux et verbaux de la phrase d'attribuer à chaque groupe nominal un numéro de rattachement si cela est possible ; enfin d'affecter aux groupes numérotés des fonctions de sujet et si un GV se trouve sans GN rattaché, d'aller lui attribuer comme sujet le groupe libre le plus proche.

Chapitre V

ATTRIBUTION

DES FONCTIONS PRIMAIRES

V.1. ATTRIBUTION DE MARQUES AUX GROUPES NOMINAUX ET AUX RELATIFS

Avant de procéder à toute recherche de fonction, nous allons agrandir de quatre positions par la gauche la place réservée au code grammatical de chacun des mots. Ce sont ces quatre positions que nous utiliserons par la suite pour marquer les différentes fonctions syntaxiques des mots ou groupes de la phrase.

V.1.1. Nécessité des marques supplémentaires

Ici encore c'est l'examen des contextes des différents groupes nominaux et des relatifs qui va être déterminant pour découvrir leur fonction. Ces tests étant très fréquents, il nous a paru nécessaire d'indiquer dans le code même de ces mots la nature de leur environnement le plus intéressant. Nous utiliserons pour ce faire, l'une des quatre positions que nous avons rajoutées à notre code.

V.1.2. Marques employées

Ce programme de traitement ne s'occupe que des marques des groupes nominaux et des relatifs. Les marques utilisées ont les significations suivantes :

valeur : blanc	pas de contexte significatif
1	groupe nominal précédé de OF
2	groupe nominal précédé d'un relatif
3	groupe nominal précédé d'une conjonction de subordinati
4	groupe nominal précédé d'une ponctuation ou débutant une phrase
5	groupe nominal précédé d'une préposition autre que OF
6	relatif précédé de la préposition OF
7	groupe nominal précédé d'un adverbe ou d'un coordonnant
A	groupe nominal précédé du mot THAT
B	groupe nominal précédé d'un mot ambigu pouvant être préposition ou conjonction.

V.2. RECHERCHE DES SUJETS ET OBJETS

V.2.1. Démarche générale

L'unité de base de cette recherche est la phrase. La recherche comporte quatre phases :

1) Un travail préparatoire sur les verbes : comptage, numérotation des verbes, examen des verbes ambigus.

2) Balayage de la phrase avec arrêt sur les verbes et recherche de leur sujet et arrêt sur les relatifs avec recherche de leur fonction.

3) Examen des catégories verbales. Comme on l'a vu au chapitre I, chaque verbe a la propriété de n'introduire que certains types de constructions dans lesquels les éléments ont une fonction bien définie. Ce sont ces différentes constructions que nous chercherons à reconnaître ici.

4) Enfin nous effectuerons un dernier balayage de la phrase portant d'une part sur les conjonctions et d'autre part sur tous les groupes ne s'étant pas vu attribuer de fonction dans l'une des trois premières phases.

V.2.2. Recherche des sujets

La recherche des sujets se fait dans la phase 2) de balayage de la phrase. Ayant trouvé un verbe conjugué non situé dans une phrase interrogative, nous pouvons chercher son sujet dans le groupe nominal le plus proche situé à sa gauche.

Ce groupe sera le sujet du verbe à condition qu'il n'ait pas l'une des marques 1, 2, 5, 7 ou B. Si le groupe possède la marque 1, nous savons alors qu'il est précédé de 'OF' ; le sujet du verbe est donc le groupe nominal précédant OF.

Si le groupe possède la marque 2), le relatif qui le précède est alors sujet. La marque 5 est traitée comme la marque 1. La marque 7 indique que le groupe nominal peut être précédé d'un adverbe ou d'un coordonnant. Le coordonnant est traité comme les marques 1 et 5. Par contre, l'adverbe et la présence sur le groupe nominal d'une marque B ne nous permettent pas d'attribuer au verbe traité un sujet absolument certain. Récapitulons ces diverses règles :

Marque du GN le plus proche à gauche	Sujet du verbe
blanc, 3, 4, 6, A	Le groupe nominal lui-même
1, 5, 7 (coordonnant)	Le groupe nominal le plus proche à gauche
2	Le groupe nominal lui-même
7 (coordonnant), B	Impossibilité d'attribuer un sujet au verbe

Il est à noter que la démarche est identique si l'on prend comme groupe de référence le groupe nominal le plus proche à droite du verbe, lorsqu'il s'agit d'une phrase interrogative. Donnons quelques exemples :

Fonctions	Marques	Groupes	Observations
Sujet	4	Equilibrated cell turnover	
		Reflects	Ier verbe
	2	The local functional demands	
Objet		Which	
Sujet	2	The environment	
		places	2ème verbe
		on	
	6	the celle population	
		.	fin de phrase

V.2.3. Recherche des objets

V.2.3.1. Introduction

L'objet peut être trouvé de façon très simple dans les relatives lorsque le relatif est lui-même objet, ou lorsqu'il est sujet et suivi d'un verbe possédant un objet.

Cela correspond aux schémas et exemples suivants :

Relatif + verbe transitif + GN
which measures 7 inches

et

Relatif + pronom sujet + verbe
which we want

Plus généralement l'objet est situé après le verbe et sa recherche procède de l'examen des catégories verbales que nous avions effectuées dès la formation du dictionnaire.

V.2.3.2. Liste des catégories verbales, schémas et exemples.

Nous présentons les diverses catégories verbales groupées par type de fonctionnement.

La zone hachurée représente le verbe et ses adverbes.

1) Le verbe est suivi d'une préposition bien définie :

Catégorie	Schémas					
04		préposition spécifique	Groupe nominal	ex. : depends ON the weather		
14		préposition spécifique	ex. : to be relied ON			
24		préposition spécifique	verbe en ING	ex. : he insisted ON going		
34		préposition spécifique	Groupe nominal	préposition quelconque	groupe nominal	ex. : I relied ON him for this business
44		préposition spécifique	groupe nominal	TO	VERBE à l'infini	ex. : I relied ON him to play

2) Schémas introduits par un adjectif ou un verbe non conjugué.

Catégorie	Schémas				
01		Adjectif ou verbe non conjugué	ex. : the boy seems ill		
23		Adjectif	THAT	ex. : IT appears Impossible THAT	
25		Adjectif	TØ	verbe à l'infinitif	ex. : IT appears impossible TO DO

N.B. : pour les catégories 23, 25 le sujet du verbe ne peut être que le pronom IT

3) Verbe précédant une conjonctive introduite par that :

Cat.	Schémas		
03		THAT	ex. : he said THAT he would comme
13		THAT	ex. : IT was decided THAT...

La catégorie 13 implique que c'est le pronom IT qui précède le verbe.

4) Le verbe est immédiatement suivi d'un groupe nominal.

Catégorie	Schémas				
02		GN ou relatif	ex. : The cat drinks milk		
06		GN	Verbe non conjugué	ex. : he hears the birds sing	
07		GN	préposition quelconque	GN ou relatif	
08		GN	GN ou relatif	ex. : I told him the news	
09		GN	adjectif verbe non conjugué	ex. : Make me live narrow	
10/30		GN	TO	verbe infinitif	ex. : I want him to come

5) Schémas commençant par TO

Catégorie	Schémas			
05		TØ	infinitif	ex. : he decided to go
15		TØ	infinitif	ex. : it was decided to go

Ici encore la catégorie 15 suppose que le sujet soit le mot IT.

V.2.3.3. Attribution des fonctions-objet

Pour chaque verbe nous avons une liste des différentes catégories verbales possibles. Elles ont été attribuées aux différentes entrées du dictionnaire en fonction des réalisations découvertes dans notre corpus.

Nous allons construire dans le texte des schémas que nous comparerons aux différents modèles que nous avons présenté ci-dessus. Un schéma étant reconnu ou une catégorie verbale étant attribuée, les fonctions des divers constituants du schéma ne posent plus de problème.

Ainsi dans les catégories 02, 06, 07, 08, 09, 10, le groupe nominal situé immédiatement après le verbe est toujours objet.

Examinons la phrase suivante :

IT appears Impossible TO DO

Le verbe possède la catégorie 25 et le schéma correspondant à cette catégorie est effectivement réalisé :

IT	sera donc sujet du verbe
Impossible	sera donc objet (au sens large)
DO	sera donc objet prépositionnel du verbe.

V.2.4. Le cas particulier des relatives

Les linguistes de la section de traduction automatique se sont très rapidement aperçus du fonctionnement particulier des relatifs, de leur petit nombre en Anglais et de leur emploi très fréquent dans la langue. Il nous a donc paru naturel de leur accorder un traitement particulier.

Nous avons effectué un relevé des différentes constructions possibles des propositions relatives. Ce relevé nous a permis de créer des modèles en nombre limité et ce sont ces modèles que nous irons confronter aux différents textes que nous aurons à analyser.

V.2.4.1. Les différentes relatives :

Quel que soit le relatif envisagé nous avons relevé trois types de propositions relatives :

type I	: le groupe du relatif est sujet du verbe
type II	: le relatif est objet direct du verbe
type III	: le relatif est précédé d'une préposition.

V.2.4.2. Les pronoms relatifs et leur fonctionnement.

Nous présentons dans un tableau à quatre colonnes, les diverses

utilisation des pronoms relatifs.

La Ière colonne comporte le nom d'un relatif

La 2ème colonne comporte son contexte

La 3ème colonne indique le type de relative que permet d'identifier le dontexte donné

Enfin dans la 4ème on indiquera si des renseignements complémentaires peuvent être tirés au sujet des groupes de ce contexte.

Relatif	Contexte	type relatif	Observations
How	How - verbe - GN	-	GN sujet du verbe
	How - adjectif - GN - verbe	-	GN sujet, adjectif : objet
However	However - Adjec. - GN - verbe		GN sujet
	However - V non c. - GN - verbe		Adjectif ou verbe non conjugué - objet
That	That - verbe	I	that sujet
What	What - verbe	I	what sujet
	What - GN - Verbe	II	what objet ; GN sujet
Where	where - verbe - GN	-	GN sujet
	where - GN - verbe	III	where circonstant - GN sujet
Which	Which - verbe	I	which sujet
	Which - GN - verbe	II	which objet ; GN sujet
Which	Prép. ≠ OF - which - verbe - GN	III	GN sujet du verbe
	GN. OF WHICH - verbe	I	GN sujet
	Verbe ing - Which - GN - verbe	-	GN est sujet du verbe
	GN - verbe ing - which - verbe	-	GN est sujet du verbe which objet du verbe en ing
Who	Who - verbe	I	Who sujet du verbe
	Who - adverbe - verbe		
Whom	Whom - GN - verbe	II	Whom objet GN sujet
Whose	Identique à what		

V.3. EXEMPLES .

Nous donnons dans ce paragraphe la liste des 21 phrases nous ayant servi de jeu d'essai pour le test des programmes d'affectation des fonctions primaires.

Nous soulignons les groupes reconnus par programmes et indiquons sous ces groupes leurs fonctions dans la phrase et leur nature (N : nominal, V : verbal, R : relatif), o correspond à une fonction non déterminée.

Phrase 1 :

The non symetrical graticules	are characterized	by	multiple axes					
N : Sujet	V		No					
which	were positionned	primarily	with	reference	to	the		
R/sujet	V			No				
distribution	to	be mapped.						
No	V							

Phrase 2 :

It	is	preferable	to	use	ink	Supplied	in	plastic tubes,	which	
N:sujet	V	A :objet		V	N:objet	V		No	R:sujet	
on	being squeezed	at	the base	will exude	from	the nozzle	the desi-			
	V		No	V		No	No			
red aucount	of	ink	of	correct fluidity.						
No		No		No						

Phrase 3 :

An original map	would be	too	long	for	a page	which	measures
N : sujet	V		AO		No	R:sujet	V
7 by 4.1 inches.							
N : objet							

Phrase 4 :

The point	which	we	want	to	emphasize	is	the following.
N : sujet	R:objet	N:obj.	V		V		N : objet

Phrase 5 :

Equilibrated cell turnover reflects the local functiomal demands
N : sujet V N : objet

which the environment places on the cell population.
R:objet N : sujet V N : objet

Phrase 6 :

Unlike the lighting of a studio photograph, in composing which
No No V R:objet

the lamps can be placed at will, the light on a subject is only at
N:sujet V No N:sujet No V

its best for one particular moment.
No No

Phrase 7 :

Certain properties are improved by reducing the arc distance which
N : sujet V V N : objet R: objet

the graticule extends from its axis.
N : sujet V No

Phrase 8 :

Hund states an additional rule that applies only to configurations
N:sujet V N : objet R:suj. V N : objet

of equivalent electrons.
No

Phrase 9 :

A third technique was used by E.R. Payne who first constructed
N : sujet V No R:suj. V

a series of dot-maps.
N : objet No

Phrase 10 :

Also, there are no experiments whose major aim is to provide
N:sujet V N: objet R:objet N :sujet V V

practice in laboratory techniques.
A :objet No

Phrase 11 :

They treat it in greater detail than is practicable in the report,
N:suj. V N:ob. No V A: objet No

in whose pages priority has necessarily be given to the mass of
N : objet N:sujet V N:objet

data which was furnished by work in the field.
No R:sujet V No No

Phrase 12 :

The flight altitude for photography depends on the purpose for
N : sujet No V N : objet

wich the pictures are made.
R:objet N : sujet V

Phrase 13 :

A study of the methods by which such topographical maps as
N:sujet No R:objet N : sujet

the various ordnance survey series or atlas map are lettered, may be
No No V V

helpful, if only to indicate the lafty standards at which to aim.
A:objet V N : objet Ro V

Phrase 14 :

A map can be drawn on which appear concentric circle to represent
N:sujet V R:objet V N : sujet V

specific distances.
N : objet

Phrase 15 :

A veritable palimpset of successive occupations shows up, -the period
N : sujet No V N :sujet

covered by which is known to range from before 2000 B.C. to
V Rc V V No

Roman times.
No

Phrase 16 :

G. Taylor devised a type of Star-diagrams based on what he considere
N :sujet V N:cbjet No V R:obj.N.suj. V

to be the four major controls of white settlement.
V N : objet No

Phrase 17 :

What finer material there is could be carried further.
N : objet N:suj. V V

Phrase 18 :

All are generalley symbolized on the basis of what is characteristic
N:suj. V No R:suj. V A : objet

for the major part of the year.
No No

Phrase 19 :

Tint-books are available, each of which has a standard number.
N : sujet V A : objet N : objet V N : objet

Phrase 20 :

A wide range of maps can be drawn, the basis of each of which
N : sujet No V N : sujet No Ro

is	the plotting	of	values	for as	many stations	as	possible.
V	N : objet		No		No		Ao

Phrase 21 :

Frequently	one	can see	a printed map,	on	the original	of	which
	N:suj.	V	N : objet		No		Ro

the lettering	has obviously been drawn	much too small.
N : sujet	V	

-:-

ANNEXES

INTRODUCTION

Les annexes que nous présentons ici ont pour but de présenter la réalisation matérielle de toutes les méthodes que nous avons développées plus haut.

En d'autres termes, nous pourrions appeler cette démarche : Reconnaissance automatique des fonctions primaires de la phrase anglaise.

Naturellement, nous avons partagé cette annexe en deux chapitres décrivant les programmes puis les résultats obtenus par ces programmes.

La programmation a été effectuée sur l'ordinateur CII. 10070 de l'Institut Universitaire de Calcul Automatique à Vandoeuvre et le langage COBOL a été pratiquement le seul langage employé.

Nous avons fait le choix de ce langage pour de multiples raisons. En premier lieu, c'est un langage universel d'où, à quelques détails près, une totale indépendance vis à vis du type d'ordinateur employé. En second lieu c'est parmi les langages universels dont nous disposons sur ce calculateur, celui qui nous semble le mieux adapté au traitement de caractères et à la gestion des fichiers. Enfin, le cadre dans lequel ces programmes ont été écrits en l'occurrence un centre de recherches linguistiques, n'est pas à négliger.

Le langage COBOL est certainement le langage pour lequel la maintenance des programmes en raison de leur simplicité de lecture, est la plus facile.

ANNEXE I :

PROGRAMMES

I.1. TRAITEMENTS PREPARATOIRES : CORPUS

I.1.1. Saisie de données - Chargement :

Comme nous l'avons signalé au chapitre I, parallèlement à nos recherches pour la traduction automatique, nous travaillons à la constitution d'un important corpus dont l'observation nous sera primordiale au fur et à mesure de l'avancement des recherches. Il est clair, en effet, que les phénomènes particuliers extrêmement rares que nous voudrons petit à petit cerner ne pourront être observés que dans un corpus extrêmement étendu. Ce nouveau corpus comporte trois parties.

1) Environ 50 articles tirés des numéros de 1971 et 1972 du magazine IMPACT.

La saisie et vérification s'est faite sur carte et sous le format suivant :

numéro d'article	col.1
numéro carte dans article	col.4
marque de carte suite ou de titre	col.9
texte	col.10 ____ col.80

Cette saisie représente un volume d'environ 50.000 cartes et comporte de l'ordre de 400.000 mots.

2) La saisie d'articles d'une série de numéros de la revue anglaise 'Géographical Magazine'.

3) Enfin la saisie d'articles tirés de numéros de la revue : 'Scientific American'.

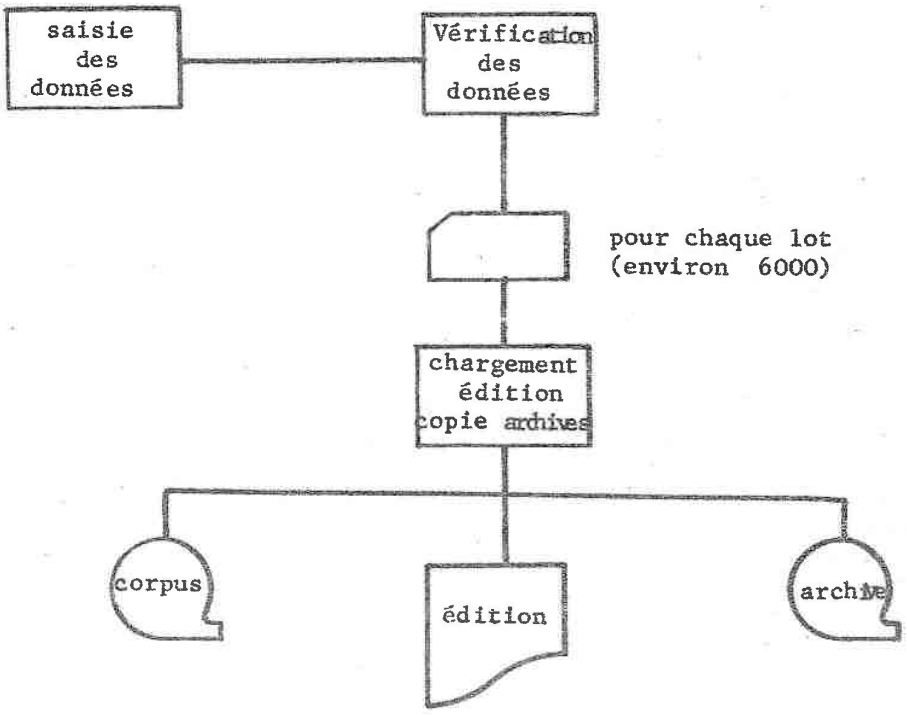
Ces deux dernières opérations sont en cours et le mode de saisie est le même que pour la revue Impact. Cependant, nous avons pour simplifier les travaux de repérage de telle ou telle phrase du texte, modifié un peu notre système de références. Les 10 premières colonnes de chaque enregistrement du fichier résultat seront donc constituées comme suit :

identification revue	col.1
numéro de page	col.4
numéro d'enregistrement dans la page	col.7
marque carte suite ou titre	col.10

Les revues 'Geographical Magazine' (du 2) recevront au fur et à mesure de leur saisie les identifications A 01, A 02 ---- tandis que les magazines 'Scientific American' (du 3) seront identifiées par les caractères B 01, B 02, etc...

Le chargement sur bande magnétique se fait dès que sont terminées la saisie et vérification d'un lot d'environ 6000 cartes (2 bacs). Il est accompagné d'un listing du texte et suivi d'une copie de toute la bande sur une bande-archive.

Chacune des trois saisies donne lieu à la création d'une bande monofichier et de sa copie. Le premier chargement d'un lot se fait dans le mode création (fichier de statut nouveau) et les chargements suivants s'effectuent dans le mode rallongement (fichier de statut MOD). C'est le même programme qui est utilisé ; il suffit en effet pour chaque mode de chargement de changer la carte d'assignation du fichier créé.

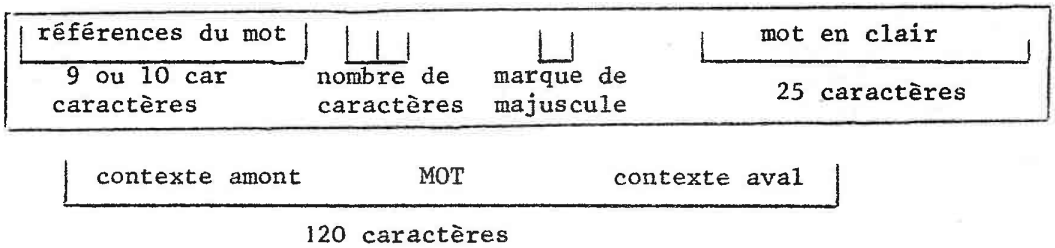


On trouvera dans l'annexe II.1. des extraits commentés de ce corpus avant tout traitement.

I.1.2. Edition des contextes

Le premier traitement que nous faisons subir à ce texte consiste à créer les contextes de chacun de ses mots, à trier le fichier obtenu dans l'ordre et éditer sous une forme adaptée aux besoins des linguistes le résultat de ce tri.

Ce programme extrêmement simple fournit cependant une quantité non négligeable de résultats intéressants. En effet, pour créer les contextes on écrit pour chaque mot du texte un enregistrement de la forme.



On voit immédiatement que la partie encadrée de l'enregistrement n'est autre que l'enregistrement que nous obtenons par un programme d'éclatement. Le programme de création des contextes fournit donc deux fichiers-résultat :

- Le fichier des mots avec leurs contextes
- Le fichier du texte éclaté.

Après avoir trié le premier de ces fichiers en utilisant la zone 'MOT en clair' comme premier argument et la zone 'Références du mot' comme argument mineur du tri, il nous est possible d'éditer les contextes sous le format suivant :

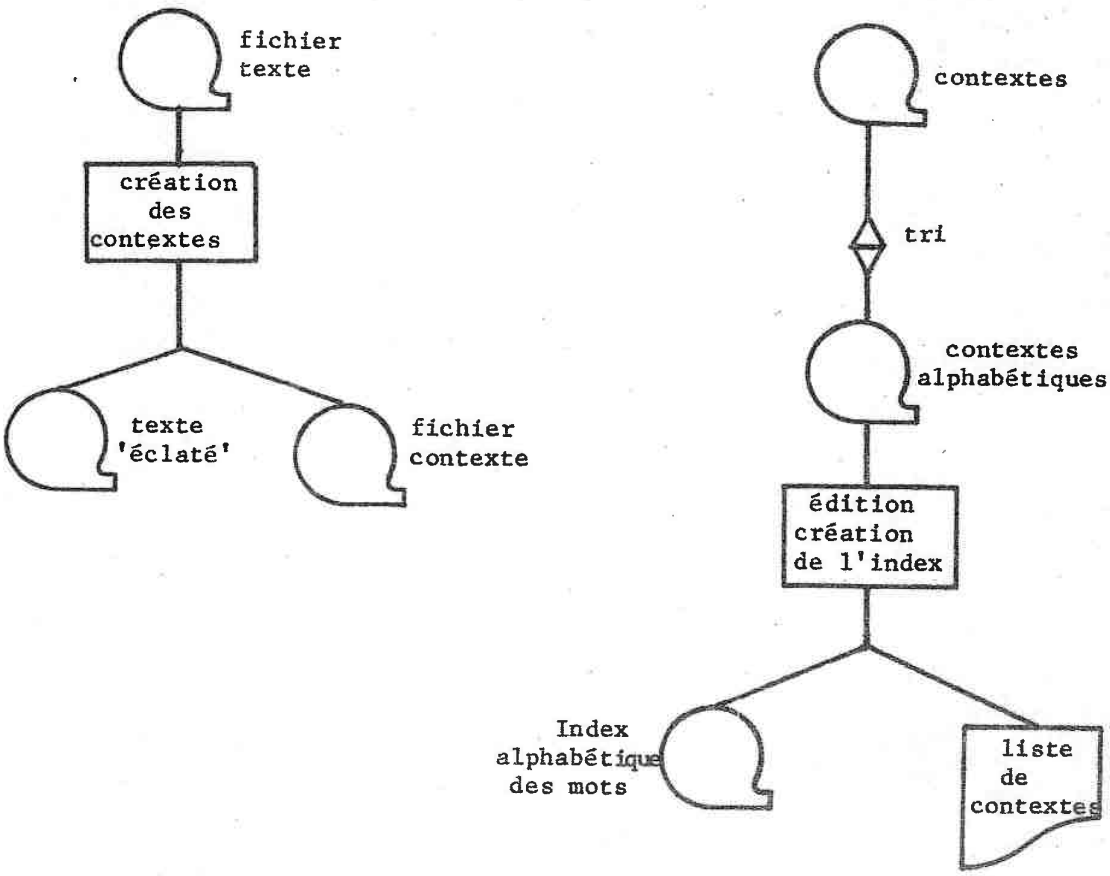
```
 ** CRAL NANCY **                                          PAGE : 1
TRADUCTION AUTOMATIQUE

                              EDITION DES CONTEXTES DE
                                   LA SERIE : IMPACT

 N°  MOTS   REFERENCES                        CONTEXTES
 1   A      B01 001 004          RELATIONS, AND AS # BASIS .
 2          B01 001 012          IN SOME PLACES  # SINGLE RESISTANT..
 3          B01 001 012          BED MAY FORM # PROEMINENT LEDGE DR
 4          B01 002 010          BEDS OUTCROPING ON # FLAT SURFACE CAR
 .
 .
 .
 .
453         B01 050 005          ROCK OUTCROPS ON # SLOPING SURFACE, THE
-----------------------------

            RECAP  :   OCCURRENCES DU MOT  A  :  453
```

L'exemple de liste de contextes que nous donnons ici comporte des contextes (3, 3) c'est-à-dire avec trois mots de part et d'autre. Pour faciliter l'observation de ces listes on remplace toujours dans la partie 'CONTEXTES' chacun des caractères du mot par le signe '# '.

Comme on le constatera à la ligne 453, les ponctuations ne comptent pas pour un mot dans la création des contextes et à la ligne 1, la création des contextes se limite au cadre de la phrase. Ainsi le Ier mot d'une phrase aura un contexte amont vide de même pour le contexte aval du dernier.

Les lignes d'édition que nous avons baptisées 'RECAP' permettent immédiatement la création du fichier index comportant les mots trouvés dans le texte avec leurs occurrences et en ordre alphabétique.

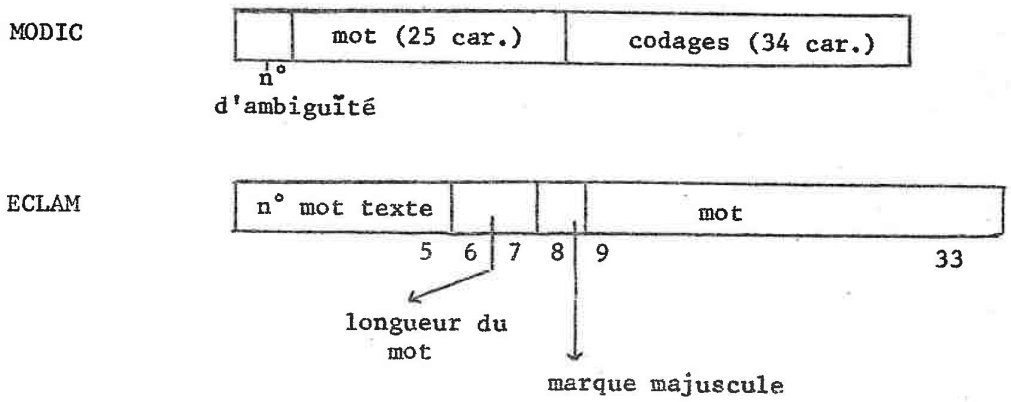
Récapitulons ces divers traitements :



I.2. CHAINE D'ANALYSE DE LA PHRASE ANGLAISE

I.2.1. Mises à jour du dictionnaire

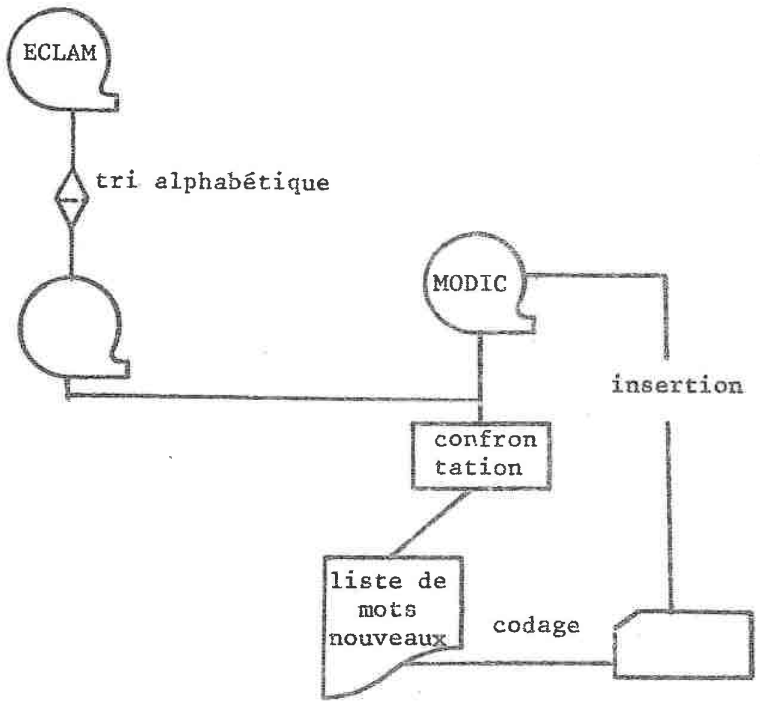
Le traitement de tout texte nouveau nécessite une mise à jour du dictionnaire. On commence en effet ce traitement par un codage des mots du texte par le dictionnaire. Ce dictionnaire étant encore relativement restreint, de nombreux mots du texte nouveau n'y figurent donc pas. Rappelons les formats des enregistrements du dictionnaire et des enregistrements du texte à traiter.



A partir de ces deux fichiers nous allons donc éditer la liste des mots ne figurant pas au dictionnaire. Cette liste sera ensuite codée par les linguistes et insérée par un programme de fusion dans le dictionnaire.

La façon la plus simple de déterminer les mots nouveaux nous a semblé être de trier le fichier 'eclam' dans l'ordre alphabétique et de lire en parallèle le dictionnaire et le fichier trié afin d'éditer tous les mots inconnus.

Un avantage de cette méthode est d'obtenir une liste d'insertion triée par ordre alphabétique d'où une mise à jour par fusion immédiate.



1. Pour chaque groupe d'enregistrement du fichier ECLAM non encore traité et concernant un même mot
   2. Ejection des mots du fichier DICO concernant des mots inférieurs dans l'ordre alphabétique.
   3. Si les derniers mots lus dans les deux fichiers ne sont pas égaux alors 4. Editer le mot du fichier ECLAM dans la liste des mots inexistants
   5. Aller à 1
6. Arrêt. fermeture des fichiers ECLAM, DICO

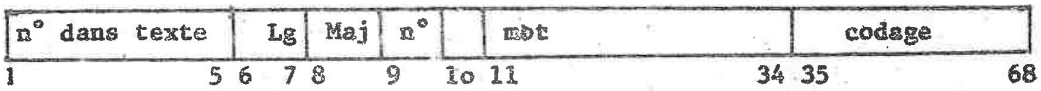
I.2.2. Préfixes - Lexies - rections.

Ce traitement consiste, sur le fichier des mots issu du codage par le dictionnaire à reconnaître les lexies et les mots à trait-d'union et à tester la présence de certaines prépositions bien définies à la suite de verbes, adjectifs ou substantifs.

Décrivons en premier lieu le format des fichiers :

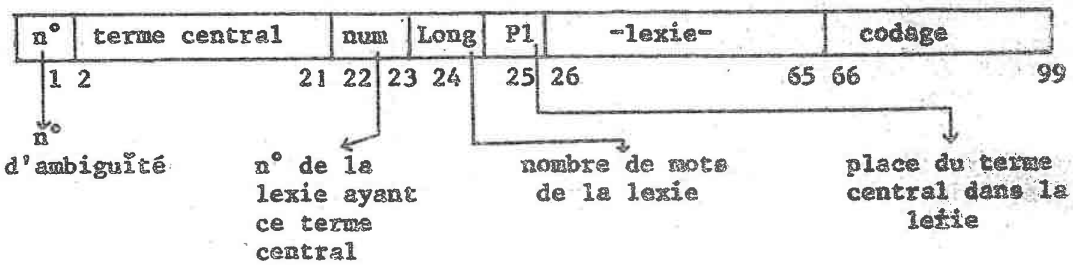
A) En entrée :

- Fichier MOT-CODE, c'est un fichier sur support magnétique, de format fixe et à accès séquentiel. Les enregistrements comportent 68 caractères :



Le n° figurant en position 9 permet de distinguer les différentes ambiguïtés de codage d'un mot donné. En colonne 10 : marque de lexie.

- Fichier LEXIQUE ; c'est un fichier disque de format fixe et à accès direct. La longueur de l'enregistrement est de 99 caractères, la clé occupe les 23 premiers.



exemple :

1	DATE	01	2	1	DATE BACK	VOA

- La lexie 'DATE BACK' comporte deux mots et le terme central DATE est en position 1.

B) En sortie :

- Fichier 'Sormaf' qui est une forme un peu plus élaborée du fichier 'MOT-CODE' puisqu'on va y trouver déjà certains groupes de mots.

- EDIT ; une édition qui est plutôt une trace puisqu'elle permet de suivre le déroulement du programme chaque regroupement donnant lieu à une ligne d'édition comportant :

  - un libellé : LEXIE RECONNUE
    ou MOT A TRAIT D'UNION

suivi du groupe effectivement construit.
D'autres libellés en cas d'erreur sont également prévus :

- marque de lexie erronée
- lexique erroné.

le premier pour un mot du texte susceptible d'engendrer une lexie et ne figurant pas au lexique, le second pour une incohérence des nombres longueur et place dans un enregistrement du lexique.

Nous avons décrit sommairement au chapitre II le fonctionnement de ce programme ; nous nous bornerons ici à présenter l'analyse modulaire de la démarche de constitution d'une lexie.

1. Pour chaque enregistrement du fichier MOT-CODE dont l'octet 10 a la valeur '*'

   2. Constituer une zone de travail comportant :
      - le caractère '1' suivi du mot, suivi du numéro 01

   3. En utilisant cette zone pour clé lire le lexique
      si la clé est invalide alors

      4. Editer le message d'erreur
         'Marque de lexie erronée' suivi de la zone MOT

      5. Allera 1

   sinon 6. Concaténer les enregistrements du fichier MOT-CODE en une expression, en utilisant les paramètres long, place trouvés au lexique.

   Si l'expression est égale à l'expression du lexique

   alors 7. Ecrire l'expression dans le fichier SORMAF

      8. Editer le libellé 'lexie reconnue'
         suivi de l'expression

      9. Ajouter 1 au Ier octet de la zone de travail créée en 2.

      10. Allera 3.

   sinon 11. ajouter 01 au numéro figurant à la fin de la zone de travail constituée en 2

      12. Allera 3.

I.2.3. Levées d'ambiguïtés.

I.2.3.1. Un processus commun :

Qu'il s'agisse des ambiguïtés conduisant au groupe verbal, de celles préparatoires au groupe nominal ou de toute autre catégorie d'ambiguïté, du point de vue de l'action des programmes sur des fichiers d'entrée, cette action est la même. Plus précisément, nous appelons cette transformation une mise à jour. C'est encore une mise à jour d'un type très particulier puisqu'elle ne comporte que des suppressions.

Lorsqu'on fait des suppressions on les fait selon des critères. Les critères peuvent être des références d'enregistrement lus sur des cartes par exemple ou être calculés ou encore être programmés.

Pour nous les critères sont programmés puisque les règles de levées d'ambiguïtés sont effectivement prévues dans les programmes. Mais les deux premières manières de présenter les critères d'élimination sont encore plus habituelles et nous en donnons ici deux exemples.

Si dans un fichier on a repéré par leurs références un certain nombre d'enregistrements à supprimer, cette suppression pourra se faire par lecture en parallèle du fichier et des références des articles à supprimer. Si par contre les enregistrements à supprimer sont ceux dont telle zone prend telle ou telle valeur, nous dirons que les critères de suppression sont calculables et nous effectuerons la suppression par calcul. Enfin, si les critères de suppression sont plus complexes, si plusieurs zones sont à tester, les critères devront être programmés. C'est le cas des levées d'ambiguïtés.

L'action des programmes sur les données et vers les résultats est donc indépendante des critères de sélection et pour nous, des ambiguïtés à lever.

En second lieu, c'est toujours le contexte qui détermine les levées d'ambiguïtés. Ce contexte est limité à la phrase. Par conséquent tous les programmes ont besoin d'avoir en mémoire toute la phrase à traiter.

Enumérons les parties communes aux divers programmes :

- Chargement de la phrase en mémoire
- Repérage des mots ambigus

- Récupération des résultats des traitements de levée d'ambiguïtés marquage des codes à éliminer.
- Recopie de la phrase en mémoire auxiliaire avec suppression des enregistrements correspondant à des codes superflus.

Si nous utilisons les notations suivantes :

Entrée : les mots du texte

Sortie : le résultat après levée

PHR : un tableau de n éléments dont chacun comporte

1 MOT (un mot ayant deux codes occupera donc deux lignes dans ce tableau)

1 CODE

1 REPERE indiquant le n° de code pour ce mot

IND1, IND2, IND3... les indices dans ce tableau d'un mot ambigu.

J, I, LEVEE des mémoires auxiliaires.

l'analyse de ce problème sera :

```
1   I = 1
2   Pour chaque enregistrement non encore traité du fichier ENTREE
      3. charger cet enregistrement dans la zone PHR (I)
      4. Incrementer I
      Si le mot n'est pas une fin de phrase
      alors 5. Allera 2
6. IND1 = 1
7. Pour chaque groupe d'éléments du tableau PHR d'indices IND1,
   IND2 ... concernant un seul mot
   Si ce mot n'a qu'une ligne dans PHR
   alors 8 incrementer IND1
   Si IND1 > I alors
         9 Allera 19
   Sinon 10. Allera 7
   Sinon 11. Concatener les divers codes de ce groupe en une chaine.
   12. Si cette chaine ne définit pas une ambiguïté à traiter
   alors 13 allera 7
   Sinon 14 Effectuer le sous programme correspondant à l'ambiguïté
            à traiter qui donnera à LEVEE une valeur choisie parmi
            0, IND1, IND2 ...
   Si LEVEE = 0 alors
         15. Allera 8
```

Sinon 16 annuler les quantités REPERE (J), pour les valeurs de J = IND1, IND2 ...
17 REPERE (LEVEE) = 1
18 Allera 8
19 Pour tous les éléments PHR (J) pour lesquels REPERE (J) ≠ 0
20 Créer avec PHR (J) un enregistrement du fichier SORTIE
21 Allera 1

Les deux programmes de levées d'ambiguïtés des verbes et des substantifs ne diffèrent sur cette analyse que par les points 12 et 14.
Ce sont ces deux points que nous développerons aux paragraphes suivants

I.2.3.2. Au sujet des verbes :

Pour effectuer les points 12, 13, 14 nous avons procédé de la manière suivante :

a) Branchement au traitement spécifique

Nous avons rangé dans un tableau AMBIG toutes les ambiguïtés verbales possibles.

Nous constituons une chaine de caractères de même structure que chaque ligne du tableau AMBIG.

Nous comparons alors la chaîne aux éléments du tableau.

Le résultat de cette comparaison est un indice I que nous utilisons comme paramètre d'un aiguillage dans le langage COBOL.

Si l'on appelle AMB1, AMB2 ... AMBn, NEGATIF respectivement les traitements des ambiguïtés 1, 2 ... n, le traitement à effectuer si l'ambiguïté trouvée dans le texte n'était pas prévue, il suffit alors d'écrire l'instruction.

```
  GO  TO  AMB1, AMB2, ... AMBn  DEPENDING  ON  I.
NEGATIF.
  .....
AMB1.
  .....
   etc...
```

Si l'ambiguïté n'était pas prévue I aura la valeur n + 1 et dans ce cas l'aiguillage se branche en séquence, c'est-à-dire au traitement NEGATIF désiré.

b) Programmation du traitement spécifique :

Le traitement spécifique d'une ambiguïté est pour nous une séquence de programme qui utilise pour donnée les indices du mot à traiter, le tableau contenant la phrase et qui fournit comme résultat dans un entier que nous appelons LEVEE :

- la valeur 0 si le contexte ne s'est pas avéré pertinent pour la levée d'ambiguïté
- la valeur de l'indice du code choisi dans le tableau PHR.

On peut donner pour exemple la levée de l'ambiguïté entre verbe passif et adjectif. C'est le cas de mots tels que INCLINED (incliné). Le tableau PHR se présente de la manière suivante :

indices	mots	codes
I1		
.		
.		
.		
I3		
I2		
I1		
IND	INCLINED	Vdp
IND1		A1
J1		
J2		

Verb-Adj.

Si COD (I1) = 'I' OU (COD (I2) = 'I' ET COD (I1) = 'K') alors LEVEE = IND1 Allera fin.

Si COD (I1) = 'X' OU COD (J1) = 'P' alors LEVEE = IND sinon LEVEE = 0.

FIN . allera prog-Gen

Autrement dit un ambigu Verbe / adjectif sera verbe s'il est précédé d'un auxiliaire (code X) ou suivi d'une préposition (code P) et sera adjectif s'il est précédé d'un introducteur ou d'un ambigu introducteur/pronom.
Dans tous les autres cas l'ambiguïté reste.

I.2.3.3. Ambiguïtés conduisant au groupe nominal.

Le schéma présenté en I.2.3.2. a) restant valable, nous nous bornerons à écrire le traitement correspondant à l'ambiguïté I/K, c'est-à-dire entre introducteur et pronom.

C'est le cas de mots tels que many, both, these, de tous les nombres également.

Le tableau PHR, de lecture de la phrase se présente de la façon suivante :

Indices	mots	codes
1		
.		
.		
.		
IND1	MANY	I
IND		K
J1		
J2		
J3		

et le traitement spécifique s'écrira

```
Intr-Pron.
   Si  COD (J1) = 'V'  ou  (COD  (J1) =  'P'  ET
   COD (J2) = 'C'  ou  'B' )  alors  LEVEE = IND
Allera fin.
   Si  COD  (J1) = 'S'  ou  'A'  ou  (COD (J1) = 'A'
   ET  COD  (J2) = 'S' )  alors  LEVEE = IND1
Allera fin.
   Si  COD (J1) =  'I'  ou  'F'  alors  LEVEE = IND1
   Sinon  LEVEE = IND.

Fin.  Allera  prog. Gen.
```

Par conséquent un ambigu introducteur/pronom sera pronom s'il est suivi d'une préposition ou d'un verbe et introducteur s'il est suivi d'un adjectif, d'un substantif ou de la séquence adjectif - substantif, ou d'un chiffre (code : F) ou d'un autre introducteur.

I.2.4. Construction et édition des groupes verbaux.

Comme la plupart des programmes constituant la chaine d'analyse de la phrase anglaise, celui de construction des groupes verbaux consiste encore en une modification de fichier.

Selon leur nature, les enregistrements du fichier d'entrée (issu de la levée d'ambiguïté sur les verbes) seront simplement recopiés s'il s'agit d'éléments non verbaux ou réunis en une chaine verbale (1 modification suivie d'un certain nombre de suppressions) lorsqu'il s'agira de termes verbaux. Les groupes verbaux se verront enfin adjoindre une zone dite d'insertion parce qu'on marquera là les codes des termes non verbaux insérés dans des groupes verbaux. Les groupes seront codés selon la nature de leurs composants suivant les modalités définies au chapitre III.

A la demande des linguistes nous avons adjoint à cette modification de fichier, une édition leur permettant de contrôler l'exactitude des regroupements et par là la qualité des algorithmes mis en oeuvre.

La structure du fichier de sortie est la suivante :

Références	Long.	MAJ	n°	MOT ou groupe	codage	insertions
1	5	6	7 8	9 10	59 60	91 92 121

Dans ce cadre on peut donner quelques exemples de représentation de groupes verbaux.

Réf.	Longueur	Maj.	n°	MOT	CODE	Insertions
1	14		1	MAM BE GROUPED	D VICP	
2	14		1	ARE CONSIDERED	D VPCP	
3	17		1	MAY BE IDENTIFIED	D VICP	
4	22		1	ARE THUS CHARACTERIZED	D VPCP	B
5	09		1	MAY OCCUR	D VICA	
6	09		1	IS BROCKEN	D VSCP	
7	02		1	IS	U VSCA	
8	14		1	ARE FLAT-LAYING	D VPCA	
9	28		1	ARE MORE EFFECTIVELY APPLIED	D VPCP	B B
10	11		1	MAY BE MADE	D VICP	
11	22		1	IS NOT TOO COMPLICATED	D VSCP	U B
12	17		1	MAY BE TRANSCATED	D VICD	
13	25		1	CAN THICKNESS BE MESURED	D VICP	SS

- Les groupes référenciés 4, 9, 11, 13 comportent des termes insérés non verbaux (soulignés) et dont les codes sont indiqués dans la colonne 'insertion' :

B adverbe, U négation, SS substantif.

I.2.5. Construction et édition des groupes nominaux.

La démarche de construction des groupes nominaux, mises à part bien entendu, les règles de formation, est identique à celle de construction des groupes verbaux.

Nous nous bornerons à donner des exemples de groupes formés et de leurs codages. On notera que la colonne d'insertion a cette fois disparu parce qu'on n'a jamais trouvé dans des groupes nominaux d'éléments spécifiquement non nominaux comme cela avait été le cas dans les verbes.

Réf	Longueur	Maj.	n°	GROUPE	CODE	
1	20		1	THE SURFACE FEATURES	D NP	
2	22		1	THE FOLLOWING CHAPTERS	U NP	
3	36		1	INTRICATE, FINE-TEXTURED DISSECTION	U NS	
4	29		1	UNCONSOLIDATED SANDY DEPONTS	U NP	
5	26		1	DIFFERENT LITHOLOGIE TYPES	U NP	
6	26		1	ARID AND SEMI-ARID REGIONS	U NP	
7	32		1	PROVISIONAL, EMPIRICAL CHARACTER	U NS	
8	22		1	THE GENRAL APPEARANCE	U NS	
9	31		1	THE COMPARATIVE CHARACTERISTICS	U NP	

I.2.6. REconnaissance des sujets/objets.

I.2.6.1. Introduction

Le programme de reconnaissance des sujets/objets travaille sur le fichier du texte avec ses groupes nominaux et verbaux et dans lequel le degré d'ambiguïté des mots est devenu très faible (inférieur à 1,1).

Ce programme étant très volumineux, nous l'avons décomposé en modules, chaque module ayant un rôle bien précis.

1) CHAR-RECOP : chargement de la phrase en mémoire
Recopie du résultat.

2) RECH : Examen des mots de la phrase avec arrêt sur les verbes, les relatifs.

3) SUJET : Recherche des sujets autour du verbe.

4) OBJET : Recherche des objets par examen des catégories grammaticales du verbe.

5) RELAT : Examen de la syntaxe des relatives.

I.2.6.2. Enchaînement des modules

1. CHAR-RECOP : chargement de la phrase à traiter en mémoire après recopie de la phrase traitée dans le fichier-résultat.
2. RECH : recherche dans la phrase d'un élément déclenchant un traitement spécifique.

Ce mot est un VERBE	O	N	N	N	N
Ce mot est un RELATIF	N	O	N	N	N
Ce mot est un RELATIF-OBJET	N	N	O	N	N
Ce mot est un RELATIF-sujet	N	N	N	O	N
Ce mot est une fin de phrase	N	N	N	N	O
3. SUJET : chercher son sujet	x				
4. OBJET : chercher son objet	x				
5. ALLERA 2	x				
6. RELAT : chercher le verbe		x	x	x	
7. ALLERA 3		x			
8. Effectuer 3			x		
9. ALLERA 2			x		
10. ALLERA 4				x	
11. ALLERA 1					x

I.2.7. Edition récapitulative.

Afin de mettre en lumière toute la démarche accomplie pour parvenir aux fonctions primaires en partant d'un texte brut sans code, nous avons un programme d'édition qui fait apparaître successivement les mots, leurs codes, leurs codes non ambigus, leur appartenance à des groupes verbaux ou nominaux, enfin les fonctions que l'on a pu attribuer à certains de ces groupes.

Ce programme travaille sur trois fichiers d'entrée qui seront lus en parallèle :

- Le fichier LEX : du texte codé avec ses lexies reconnues et les mots à traits d'union.
- Le fichier GV : du texte après regroupement verbal
- Le fichier SYN : après regroupement nominal et affectation des fonctions.

Le résultat (dont on trouvera un exemple dans l'annexe II) est une édition comportant trois colonnes.

De gauche à droite :

- Les codes définitifs des groupes avec leurs fonctions en clair : verbe, n° ..., sujet du verbe, objet du verbe.
- Une colonne indiquant les regroupements nominaux
- Une colonne indiquant les regroupements verbaux et la liste des mots dans l'ordre du texte avec leurs codes issus du dictionnaire.

Ainsi, dans la phrase :

'The non-symetrical graticules are characterized by multiple axes'

1) Le groupe nominal "The non-symetrical graticules" sera édité comme suit :

Sujet du V, 1	U	NP	***	I	**	I	The
			*				
			*	A	**	A	non-symetrical
			*				
			*	SP	**	SP	graticules
			*				

groupe nominal sujet
du verbe n° 1 de la phrase

2) Le groupe verbal 'are characterized'

Verbe 1	D VPCP	*** D VICP	** 1XS	ARE
			* 1VSA	
			*	
groupe verbal 1			* VDA	Characterized
de la phrase			* VDP	
			*	

3) et le mot BY qui du début à la fin n'a subi ni levée d'ambiguïté ni insertion dans un groupe :

UP	*** PQ	** PQ	BY

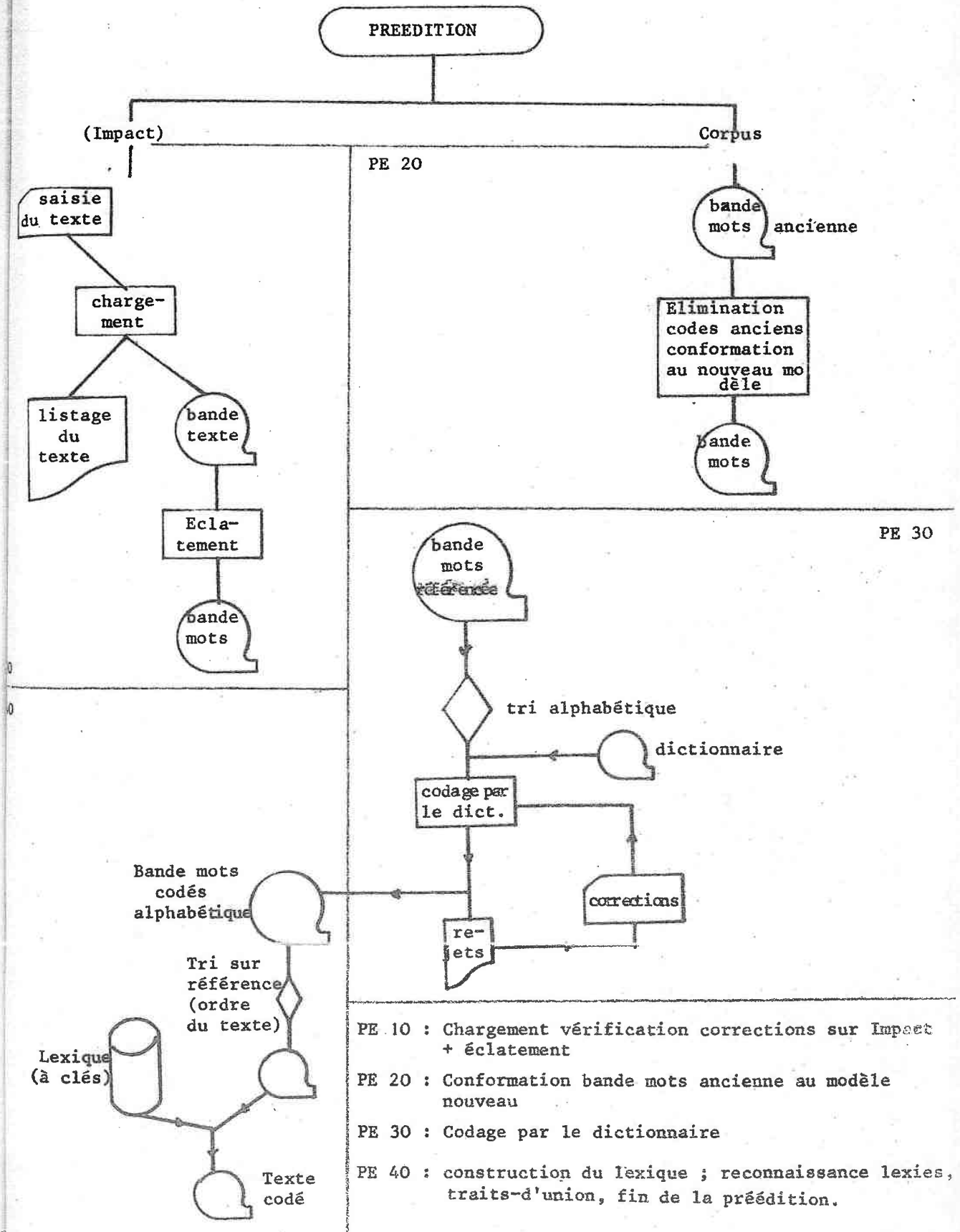
d'autre part il y a retour en haut de page pour tout début de phrase qui est spécifié par le libellé :

DEBUT DE PHRASE nnn. (lib ...)

nnn : n° de la phrase dans le texte

lib.. un libellé lu sur une carte paramètre identifiant le texte traité.

I.2.8. Vue d'ensemble du traitement

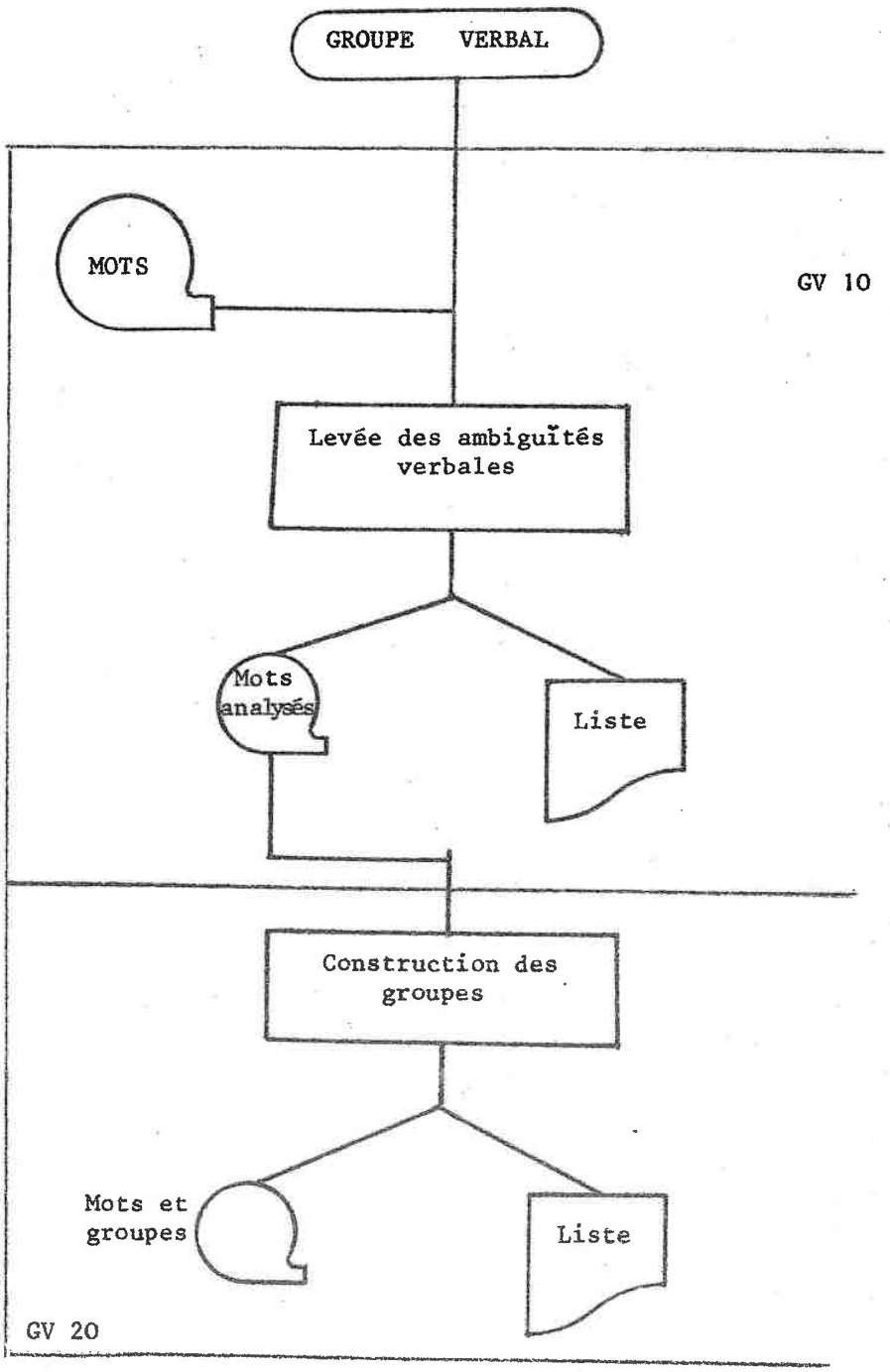


PE 10 : Chargement vérification corrections sur Impact + éclatement

PE 20 : Conformation bande mots ancienne au modèle nouveau

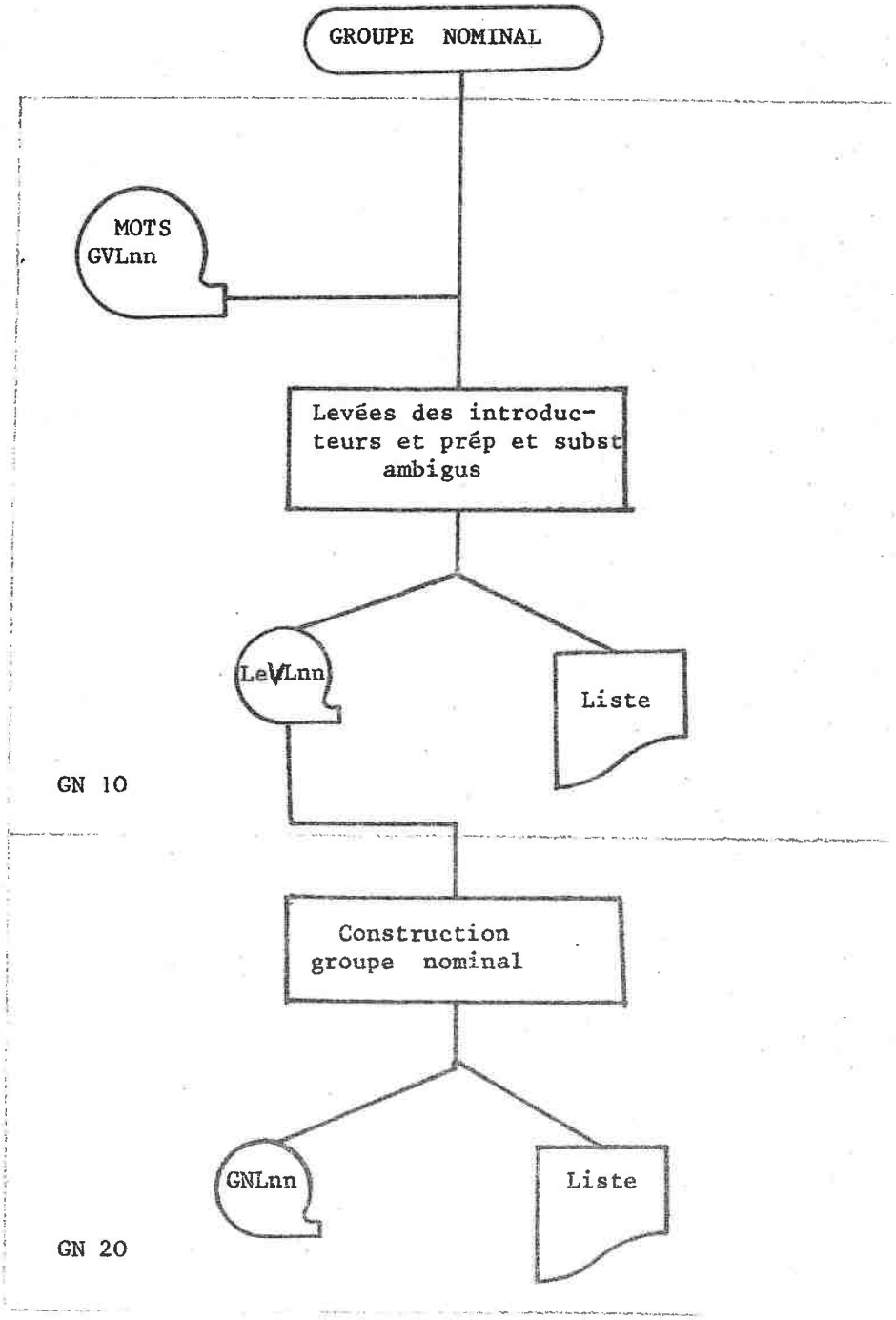
PE 30 : Codage par le dictionnaire

PE 40 : construction du lexique ; reconnaissance lexies, traits-d'union, fin de la préédition.



GV 10 : programme de levée des ambiguïtés sur tous les verbes

GV 20 : regroupement des verbes en groupes verbaux + édition des résultats.



GN 10 : Levée des ambiguïtés préparatoires au GN

GN 20 : Reconnaissance et construction des GN.

ANNEXE II :

RESULTATS

Les annexes que nous présentons dans les pages suivantes portent sur les résultats obtenus à partir des divers modules composant la chaîne d'analyse automatique de la phrase anglaise.

Les titres et contenus de ces annexes sont les suivants :

Annexe II.2. : Un exemple de la saisie de donnée du magazine 'Scientific American'.

Annexe II.2. : Des extraits des dictionnaires.

II.2.1. : Les mots commençant par la lettre N et appartenant au dictionnaire des mots.

II.2.2. : Les lexies dont le terme central commence par l'une des lettres C, D, E ou F.

Annexe II.3. : Une image de la démarche de levée d'ambiguïtés verbales où l'on voit apparaître successivement le fichier d'entrée, les messages émis par le programme de levée d'ambiguïtés puis le fichier de sortie.

Annexe II.4. : Un exemple de l'édition qui est associée au regroupement verbal.

Annexe II.5. : Les messages émis par le programme de reconnaissance des sujets/objets ainsi que les codes affectés au texte par ce programme.

Annexe II.6. : Un exemple de l'édition récapitulative de tout le traitement.

II.1. SAISIE DE DONNEES.

```
B01052001@*TECHNØLØGY AND *ECØNØMIC *DEVELØPMENT BY *ASA
B01052002@*BRIGGS
B01052003 PRESENTING AN ISSUE DEVØTED TØ THE PRØBLEM ØF HØW
B01052004 NATIØNS CAN ATTAIN A STATE ØF SELF - SUSTAINING GR
B01052005-ØWTH .
B01052006 THIS ARTICLE ØUTLINES THE HISTØRY ØF DEVELØPMENT A
B01052007-ND ØF THE DIVISIØN ØF NATIØNS INTØ " RICH " AND "
B01052008 PØØR " .
B01052009 A CIRCUMSTANCE NEW TØ HISTØRY STRETCHES THE TENSIØ
B01052010-NS ØF CØNTEMPØRARY WØRLD PØLITICS .
B01052011 THIS IS THE WIDESPREAD AWARENESS ØF THE DIVISIØN Ø
B01052012-F NATIØNS INTØ TWØ CLASSES : " DEVELØPED " AND " U
B01052013-NDERDEVELØPED " IN THE PARLANCE ØF THE DAY ØR , IN
B01052014 PLAINER WØRDS , RICH AND PØØR .
B01052015 THE CØNTØUR LINES ØF INTERNATIØNAL ECØNØMIC INEQUA
B01052016-LITY ARE EASILY DRAWN .
B01052017 TØ THE CLASS ØF THE RICH BELØNG THE NATIØNS ØF NØR
B01052018-THWESTERN *EURØPE AND THØSE ELSEWHERE IN THE *TEMP
B01052019-ERATE *ZØNES THAT WERE SETTLED AND ØRGANIZED BY PE
B01052020-ØPLE ØF THE SAME STØCK : ØHE *U.S. , *CANADA , *AU
B01052021-STRALIA AND *NEW *ZEALAND .
B01052022 ØNE NØN - *EURØPEAN NATIØN -- *JAPAN -- SHØULD AL
B01052023-SØ BE CØUNTED IN THE GRØUP , AND RECENTLY ANØTHER
B01052024 *EURØPEAN NATIØN , THE *U.S.S.R. , HAS JØINED IT .
B01052025 THESE NATIØNS , CØNSTITUTING LESS THAN A THIRD ØF
B01052026 THE HUMAN PØPULATIØN , PRØDUCE AND CØNSUME MØRE TH
B01052027-AN TWØ - THIRDS ØF THE WØRLD ES GØØDS .
B01052028 THEIR ØUTPUT IS INCREASING MØRE RAPIDLY THAN THEIR
B01052029 PØPULATIØN , AND THEY BØAST RISING INCØMES PER CAP
B01052030-ITA .
B01052031 INCØME PER CAPITA HARDLY SERVES AS A MEASURE ØF TH
B01052032-E PØSITIØN ØF THE NATIØNS ØF THE PØØR .
B01052033 THE ØVERWHELMING MAJØRITY ØF THEIR PØPULATIØNS ARE
B01052034 ØCCUPIED IN SUBSISTENCE AGRICULTURE AND LIVE ALMØS
B01052035-T ENTIRELY ØUTSIDE THE MØNETARY SYSTEMS ØF THEIR M
B01052036-EAGER ECØNØMIES .
B01052037 FØR WHAT THE ECØNØMIC INDICES ARE WØRTH , THEY SHØ
B01052038-W THAT BETWEEN THE PØØREST 1.5 BILLIØN PEØPLE -- T
B01052039-HE BØTTØM HALF ØF THE HUMAN PØPULATIØN -- AND THE
B01052040 AVERAGE STANDARD ØF LIVING PREVAILING IN THE RICH
B01052041 CØUNTRIES THE DISPARITY IS ØN THE ØRDER ØF ØNE TØ
B01052042 10 .
B01052043 MØRE SIGNIFICANTLY , THE INDICES SHØW THAT THE DIS
B01052044-PARITY BETWEEN THE TWØ CLASSES ØF NATIØNS IS WIDEN
B01052045-ING .
B01052046 PØVERTY IS NØT , ØF CØURSE , A NEW CØNDITIØN IN HU
B01052047-MAN AFFAIRS .
B01052048 SØME ØF THE PØØR NATIØNS WERE ØNCE WØRLD PØWERS AN
B01052049-D WERE HELD TØ BE RICH AS WELL AS PØWERFUL .
B01052050 BUT EVEN THØUGH THEY HAVE BEEN PLACED AT A DISADVA
B01052051-NTAGE IN RECENT YEARS BY THE UNFAVØRABLE TERMS ØF
B01052052 THEIR RELATIØNS WITH THE RICH NATIØNS , THE SITUAT
B01052053-IØN IN WHICH THEIR PEØPLES LIVE IS NØT MUCH WØRSE
B01052054 THAN BEFØRE .
B01052055 THE RICH NATIØN IS THE NØVELTY , AND THE DEVELØPME
B01052056-NT THAT MAKES ENTIRE NATIØNS RICH IS ITSELF THE PI
B01052057-VØTAL DEVELØPMENT ØF MØDERN HISTØRY .
B01052058 TØ UNDERSTAND THE INCREASING ECØNØMIC INEQUALITY Ø
B01052059-F NATIØNS ØNE MUST LØØK ØUTSIDE THE BØUNDARIES ØF
B01052060 ECØNØMIC THEØRY .
B01052061 IN THE SEARCH FØR THE CAUSES , ANTECEDENTS AND " P
B01052062-RECØNDITIØNS " ØF DEVELØPMENT IT IS NECESSARY TØ T
B01052063-URN TØ HISTØRY , AND THE HISTØRIAN HAS HIS CHØICE
B01052064 ØF STARTING PØINTS .
```

NATIONAL	A	*00	
NATURAL	A	00	
NATURE	A	00	
NATURE	SS		
NAVAL	A	00	
NAVY	SS		
NEAR	P	*	
NEAR	B		
NEARBY	B		
NEARER	1A	0400	PC
NEARLY	B1		
NEBULOUS	A	00	
NECESSARILY	B		
NECESSARY	A	04232520	PF
NEED	YO		
NEED	VOA	051102	
NEED	SS	040500	PF
NEEDED	VDP	0400	PF
NEGATIVE	SS		
NEGATIVELY	B		
NEVERTHELESS	B		
NEW	A	0400	PC
NEXT	A1	*00	
2NEXT	B		
NICE	A	00	
NINE	IPC		
2NINE	KPC		
NO	IG	*	
NON	1G		
NORMAL	A	04	PC
NORMALLY	B		
NORTH	A1	00	
2NORTH	SS		
3NORTH	B		
NORTHWARD	A1	00	
NOT	U		
NOTED	VDA	0302	
2NOTED	VDP	1300	
NOTHING	KS		
NOTICEABLE	A	00	
NOW	B		
NUMBER	SS		
NUMBERS	VSA	02	
2NUMBERS	SP		
NUMEROUS	A	00	
NUTRIENT	A1	00	
2NUTRIENT	S1	00	

```
ALLED                0122SO CALLED                                  A    00
ALLED                0221CALLED UP                                  VDP  00
ARRIED               0121CARRIED OUT                                VDA  02
ARRY                 0121CARRY OUT                                  VDA  02
ARRYING              0121CARRYING OUT                               SS
ENT                  0122PER CENT                                   A1
ENT.                 0122PER CENT.                                  A1
LOSED                0121CLOSED DOWN                                VDP  00
OPRINUS              0121COPRINUS LAGOPUS                           PSS
OURSE                0122OF COURSE                                  B
OVERED               0121COVERED IN                                 VDP  00
ROP                  0121CROP OUT                                   VDA  00
ULTURING             0133SUB - CULTURING                            SS


ATE                  0121DATE BACK                                  VDA  04
          PC     **  C A R T E   S U I T E      **
ATE                  0233PRE - DATE                                 VDA  02
ATE                  0333POST - DATE                                VDA  02
ATES                 0133POST - DATES                               VSA  02
IRECT                0121DIRECT FROM                              D PI
ISTANCE              0121DISTANCE APART                             SS
RAWING               0121DRAWING UP                                 VDA  02
RAWING               0221DRAWING INSTRUMENT                         SS
RAWING               0321DRAWING INSTRUMENTS                        SP
RAWING               0421DRAWING TIME                               SS
RAWING               0522COLOUR DRAWING                             SS
RAWING               0622LAND DRAWING                               SS
RAWING               0722LANDSCAPE DRAWING                          SS
RAWING               0822LINE DRAWING                               SS


EACH                 0121EACH OTHER                                   K
EVEN                 0121EVEN IF                                  D C
EVEN                 0221EVEN THOUGH                              DTC
EVEN                 0321EVEN WHEN                                D C
EXAMPLE              0122FOR EXAMPLE                                  B
EXCEPT               0121EXCEPT FOR                               D PP
EXCEPT               0221EXCEPT THAT                              D CT


FACT                 0122IN FACT                                      B
FAR                  0143IN SO FAR AS                             DTCD
FAR                  0232AS FAR AS                                D CD
FAR                  0332SO FAR AS                                D CD
FAR                  0421FAR APART                                  A2
FAR                  0521FAR AWAY                                   A2
FAR                  0621FAR BETWEEN                                A2
FAR                  0721FAR BACK                                   A2
FAR                  0822THUS FAR                                   B
FAR                  0921FAR INTO                                   B
FILLED               0121FILLED UP                                  VDA  0400
          PH     **  C A R T E   S U I T E      **
FIRST                0122THE FIRST                                    I
FORMER               0122THE FORMER                               SI
FURTHER              0121FURTHER AWAY                                 B
```

II.3. ETAT DU FICHIER D'ENTREE


```
0385  1 THE                    I
0386  1 NEED                   YO      054102
0386  2                        VOA     040500
0386  3                        SS
0387  1 FOR                    PEN
0387  2                        CF
0388  1 MORE                   1I
0388  2                        1K
0388  3                        1B
0389  1 EXTENSIVE              A       00
0390  1 TRAINING               SS
0391  1 IN                     PG
0392  1 THE                    I
0393  1 USE                    VOA     02
0393  2                        SS
0394  1 OF                   U PA
0395  1 AERIAL                 A1      00
0396  1 PHOTOS                 SP
0397  1 HAS                  2XS
0397  2                      2VSA    020609
0398  1 BEEN                 1XN
0398  2                      1VNA    040123250200
0399  1 EMPHASIZED           VDP     1300
0400  1 BY                   PG
0401  1 MILITARY             A       00
0402  1 AND                  E1
0403  1 NAVAL                A       00
0404  1 OFFICIALS            SP
```

LEVEE DES AMBIGUITES VERBALES

MOT : 386 AMBIGUITE : YO / VOA / SS LEVEE : SS

MOT : 393 AMBIGUITE : VOA / SS LEVEE : SS

MOT : 397 AMBIGUITE : 2XS / 2VSA LEVEE : 2XS

MOT : 398 AMBIGUITE : 1XN / VNA LEVEE : 1XN

ETAT DU FICHIER DE SORTIE

```
0385  1 THE                    I
0386  1 NEED                   SS
0387  1 FOR                    PEN
0387  2                        CF
0388  1 MORE                   1I
0388  2                        1K
0388  3                        1B
0389  1 EXTENSIVE              A       00
0390  1 TRAINING               SS
0391  1 IN                     PG
0392  1 THE                    I
0393  1 USE                    SS
0394  1 OF                   U PA
0395  1 AERIAL                 A1      00
0396  1 PHOTOS                 SP
0397  1 HAS                  2XS
0398  1 BEEN                 1XN
0399  1 EMPHASIZED           VDP     1300
0400  1 BY                   PG
0401  1 MILITARY             A       00
0402  1 AND                  E1
0403  1 NAVAL                A       00
0404  1 OFFICIALS            SP
```

REGROUPEMENT VERBAL

LONGUEUR	AMB	MOT EN CLAIR	CODE
		TRAITEMENT DE LA PHRASE NO : 15	
3	1	THE	I
4	1	NEED	SS
3	1	FOR	PEN
3	2	FOR	CF
4	1	MORE	1I
4	2	MORE	1K
4	3	MORE	1B
9	1	EXTENSIVE	A
8	1	TRAINING	SS
2	1	IN	PG
3	1	THE	I
3	1	USE	SS
2	1	OF	U PA
6	1	AERIAL	A1
6	1	PHOTOS	SP
12	1	HAS BEEN EMPHASIZED	D2VSCP
2	1	BY	PG
9	1	MILITARY	A
3	1	AND	E1
5	1	NAVAL	A
9	1	OFFICIALS	SP
1	1	,	
3	1	AND	E1
5	1	WITHIN	P
3	1	THE	I
4	1	YEAR	SS
4	1	MANY	IP
4	2	MANY	KP
8	1	COLLEGES	SP
3	1	AND	E1
12	1	UNIVERSITIES	SP
15	1	HAVE INTRODUCED	D2VICA
15	2	HAVE INTRODUCED	D2VINA
3	1	NEW	A
7	1	COURSES	SP
3	1	FOR	PF
3	2	FOR	CF
4	1	THAT	U CT
4	2	THAT	IS
4	3	THAT	KS
4	4	THAT	LT
7	1	PURPOSE	SS
1	1	.	

```
II.5. TRAITEMENT PHRASE:  +0000000001
      TRAITEMENT SUJET OBJET
II.5. CATEGORIE -00- PAS DE TRAIT                                    122
      RECHERCHE DES SUJETS
      *** REL5 *** INTRODUCTEUR ... WHICH
       SUJET SEUL TROUVE

      CATEGORIE -00- PAS DE TRAIT
      TRAITEMENT CT5
      CATEGORIES NON REALISEES OU IMPREVUES
      RECHERCHE DES SUJETS
      RECOPIE PHRASE NB MOTS :  +0000000014
      +0000000001 1 THE NON-SYMMETRICAL GRATICULES          11 4U NP
      +0000000002 1 ARE CHARACTERIZED                        1   D VPCP  00
      +0000000003 1 BY                                          U PG
      +0000000004 1 MULTIPLE AXES                            6U NP
      +0000000005 1 WHICH                                   21   LN
      +0000000006 1 WERE POSITIONED                         2L  D VPCP  00
      +0000000007 1 PRIMARILY                                   B
      +0000000008 1 WITH                                        U PH
      +0000000009 1 REFERENCE                                6U NS
      +0000000010 1 TO                                          U PCN
      +0000000011 1 THE DISTRIBUTION                         6U NS
      +0000000012 1 TO                                          U PC
      +0000000013 1 BE MAPPED                               3   D VINP  00
      +0000000014 1 .
      TRAITEMENT PHRASE:  +0000000002
      TRAITEMENT SUJET OBJET
      TRAITEMENT CT1
      TRAITEMENT CT3
            AFFECTATION DE L OBJET *****
      TRAITEMENT CT3
      CATEGORIE :  25 IDENTIFIEE
      RECHERCHE DES SUJETS
      TRAITEMENT CT5
            AFFECTATION DE L OBJET *****
      CATEGORIE :  02 IDENTIFIEE
      RECHERCHE DES SUJETS
      CATEGORIE -00- PAS DE TRAIT
      RECHERCHE DES SUJETS
      *** REL5 *** INTRODUCTEUR ... WHICH
       SUJET SEUL TROUVE

      TRAITEMENT CT5
      CATEGORIES NON REALISEES OU IMPREVUES
      RECOPIE PHRASE NB MOTS :  +0000000024
      +0000000001 1 IT                                      11 4U N2
      +0000000002 1 IS                                       1   UIVSCA  2
      +0000000003 1 PREFERABLE                                  12  A     0
      +0000000004 1 TO                                          13U PC
      +0000000005 1 USE                                         2   U VINA  0
      +0000000006 1 INK                                         22U NS
      +0000000007 1 SUPPLIED                                    3   U VINP  0
      +0000000008 1 IN                                          U PG
      +0000000009 1 PLASTIC TUBES                            6U NP
      +0000000010 1 ,                                        5  7  LN
      +0000000011 1 WHICH                                       U PJ
      +0000000012 1 AN                                          4   D VGNP  0
      +0000000013 1 BEING SQUEEZED                              U PN
      +0000000014 1 AT                                       6U NS
      +0000000015 1 THE BASE                                 5L  D VICA  0
      +0000000016 1 WILL EXUDE                                  U P1
      +0000000017 1 FROM                                     6U NS
      +0000000018 1 THE NOZZLE                                  U NS
      +0000000019 1 THE DESIRED AMOUNT                          U PA
      +0000000020 1 OF                                       5U NS
      +0000000021 1 INK                                         U PA
      +0000000022 1 OF                                       5U NS
      +0000000023 1 CORRECT FLUIDITY
```

II.6. Edition récapitulative.

```
DEBUT DE PHRASE 0001. (LIVRE 001)

SUJ. DU V. 1 U NP          ***  I      **  I      THE
                           *
                           *  A      **  A      NON-SYMETRICAL
                           *
                           *  SP     **  SP     GRATICULES
                           *

VERBE   V. 1 D VPCP        ***D VICP   **  1XS    ARE
                                       *  1VSA
                                       *
                                       *  VDA    CHARACTERIZED
                                       *  VDP
                                       *

          U PQ             ***  PQ     **  PQ     BY

          U NP             ***  A      **  A      MULTIPLE
                           *
                           *  SP     **  SP     AXES
                           *

SUJ. DU V. 2   LN          ***  LN     **  LN     WHICH

VERBE REL. 2   VPCP        ***D VICP   **  1XP    WERE
                                       *  1VPA
                                       *
                                       *  VDA    POSITIONED
                                       *  VDP
                                       *

            B              ***  B      **  B      PRIMARILY

          U PM             ***  PM     **  PM     WITH

          U NS             ***  SS     **  SS     REFERENCE

          U PCN            ***U PC     ** U PC    TO

          U  NS            ***  I      **  I      THE
                           *
                           *  SS     **  SS     DISTRIBUTION
                           *

          U PC             ***U PC     ** U PC    TO

VERBE   V. 3 D VINP        ***D VINP   **  1XA    BE
                                       *  1VDA
                                       *
                                       *  VDA    MAPPED
                                       *  VDP
                                       *

FIN DE PHRASE 0001                                *
```

DEBUT DE PHRASE 0002. (LIVRE 001)

				***			**	.
SUJ. DU V. 1		NZ		***	KZ		** KZ	IT
VERBE V. 1	U1VSCA			***U	VICA		** 1XS	IS
							** 1VSA	
OBJ. DU V. 1		A		***	A		** A	PREFERABL
1	U PC			***U	PC		** U PC	TO
VERBE V. 2	U VINA			***	U VI		** VDA	USE
							** SS	
OBJ. DU V. 2	U NS			***	SS		** SS	INK
VERBE V. 3	U VINP			***U	VINP		** VDA	SUPPLIED
							** VDP	
	U PG			***	PG		** PG	IN
	U NP			***	A	/ SS	** A	PLASTIC
				*			** SS	
				*				
				*	SP		** SP	TUBES
				*				
				***			**	,
SUJ. DU V. 5		LN		***	LN		** LN	WHICH
	U PJ			***	PJ		** PJ	ON
VERBE V. 4	D VGNP			***D	VGNP		** 1XG	BEING
							* 1VGA	
							*	
							* VDA	SQUEEZED
							* VDP	
							*	
	U PN			***	PN		** PN	AT
	U NS			***	I		** I	THE
				*				
				*	SS		** SS	BASE
				*				
VERBE REL. 5	D VINA			***D	VICA		** YM	WILL
							* VDA	
							*	
							* VDA	EXUDE
							*	
	U PI			***	PN		** PI	FROM
	U NS			***	I		** I	THE
				*				
				*	SS		** SS	NOZZLE
				*				

```
U NS        *** I     ** I    THE
              *
              * A     ** VDA  DESIRED
              *       ** VDP
              *       ** A
              *
              * SS    ** SS   AMOUNT
              *
U PA        *** PA    ** PA   OF
U NS        *** SS    ** SS   INK
U PA        *** PA    ** PA   OF
U NS        *** A     ** A    CORRECT
              *
              * SS    ** SS   FLUIDITY
              *
FIN DE PHRASE 0002
```

```
   DEBUT DE PHRASE 0003. (LIVRE 001)

                                ***                  **
SUJ. DU V. 1 U NS               ***  IS              **  IS      AN
                                 *
                                 *  A               **  A       ORIGINAL
                                 *
                                 *  SS              **  SS      MAP
                                 *

VERBE   V. 1 D1VICA             ***D VICA           **  VM      WOULD
                                                    *   VMA
                                                    *
                                                    *   1X8     BE
                                                    *   1VBA
                                                    *

          B                     ***  B              **  B       TBB
OBJ. DU V. 1   A                ***  A              **  A       LONG
             U PF               ***  PF   /   CF    **  PF      FOR
                                                    **  CF

                                ***  I              **  I       A
                                 *
                                 *  SS              **  SS      PAGE
                                 *

SUJ. DU V. 2   LN               ***  LN             **  LN      WHICH

VERBE REL. 2 U VSCA             ***U VSCA /   SP    **  VBA     MEASURES
                                                    **  SS

            NF                  ***  F              **  F       7
          U PQ                  ***  PQ             **  PQ      BY
          U NP                  ***  F              **  F       4.1
                                 *
                                 *  SP              **  SP      INCHES
                                 *

   FIN DE PHRASE 0003
```

```
DEBUT DE PHRASE 0004. (LIVRE 001)
                                      ***  .               **           .
            U NS                      ***  I               **    I      THE
                                        *
                                        *  SS              **    VØA    PØINT
                                        *                  **    SS
                                        *

BJ. DU V. 1    LN                     ***  LN              **    LN     WHICH
UJ. DU V. 1   JNP                     *** JKP              **   JKP     WE
VERBE REL. 1 U VICA                   ***U VICA            **    VØA    WANT
             U PC                     ***U PC              ** U PC      TØ
VERBE    V. 2 U VINA                  ***U VINA            **    VØA    EMPHASIZE
VERBE    V. 3 U1VSCA                  ***U VSCA            **   1XS     IS
                                                           **   1VSA
ØBJ. DU V. 3 U NS                     ***  I               **    I      THE
                                        *
                                        *  SS              **    VGA    FØLLØWING
                                        *                  **    A1
                                        *                  **    SS
                                        *                  **    P
                                        *
  FIN DE PHRASE 0004                                                    .
```

CONCLUSION

C O N C L U S I O N

Les conclusions que l'on peut formuler à la suite de ce travail portent d'une part sur l'amélioration de ce qui est déjà fait et d'autre part sur les horizons ouverts par les méthodes mises en oeuvre.

On a vu comment on pouvait résoudre les problèmes posés par la levée d'ambiguïté et comment cette question pouvait se ramener à une mise à jour de fichier au moyen de critères de sélection.

La méthode effectivement employée par nous et qui commence à donner des résultats est celle qui consiste à programmer ces critères.

Une amélioration sensible des temps de traitement pourrait être, à notre sens, obtenue en posant, dans les divers sous-programmes de levée d'ambiguïtés, les tests dans un ordre bien défini : celui de la plus forte probabilité d'apparition.

Cette probabilité peut être obtenue par une étude statistique d'un choix de textes codés permettant de calculer les fréquences relatives des différentes chaines de codes correspondant à deux, trois ou quatre mots consécutifs.

Il serait également possible de changer totalement le principe de levée des ambiguïtés en interprétant les critères de sélection par un programme au lieu de les programmer.

Enfin, le dictionnaire pourrait être notablement allégé par une étude de morphologie sur les mots. Nous pensons ici aux verbes réguliers qui n'auraient plus qu'une entrée (infinitif) au lieu de quatre (infinitif, forme en -s, forme en -ED, forme en -ING) et aux substantifs qui ne figureraient plus qu'au singulier.

Néanmoins, cette amélioration ne saurait se concevoir que pour le code grammatical puisque les codes syntaxiques des verbes dépendent de la forme à laquelle ils sont employés.

Pour en terminer avec ce premier volet de conclusion, il nous faut signaler certaines erreurs relevées au fur et à mesure des traitements effectués sur les différents textes composant notre corpus. Ces erreurs, provenant soit de cas non prévus (parce que très rares) soit d'erreurs mineures de programmation, font l'objet d'actives corrections. Elles ont

l'avantage de permettre d'affiner l'analyse linguistique et cette méthode n'est pas sans analogie avec celle des approximations successives. Le pourcentage d'erreur au fur et à mesure de l'extension à de nouveaux textes va donc en décroissant.

D'autre part, il ne faut pas négliger le but final qui est, un jour, de traduire des textes écrits en Anglais scientifique. Les problèmes rencontrés dans l'analyse de la phrase, conjugués à cet objectif permettent d'envisager les questions linguistiques de concordance de temps et d'organisation du dictionnaire bilingue du nombre de synonymes français à y faire figurer.

-:-

BIBLIOGRAPHIE

/1/ : *Etudes sur l'analyse de la phrase allemande dans le système context-free* (Thèse de 3ème cycle, E. GALICHET, Grenoble 1968).

/2/ : *Exploitation d'un corpus d'anglais scientifique écrit* (Cahier du C.R.A.L. n° 15, 1970).

/3/ : *Etude de la discontinuité dans le syntagme verbal en Anglais scientifique écrit* (Thèse de 3ème cycle, J. LECOMTE, Nancy, 1971).

/4/ : *Rapport de la section de Traduction Automatique*, dans le rapport d'activité du C.R.A.L. 1973.

/5/ : CHOMSKY - MILLER : *Analyse formelle des langues naturelles* (Mouton, Paris - La Haye).

/6/ : *Génération automatique des verbes français* (D.E.A. par C. Sanchez, Nancy 1973).

/7/ : BLOIS et BUYDENS : *Règles catégorielles de levées d'ambiguïtés pour l'analyse syntaxique automatique du français.* (Revue : Problèmes de la Traduction Automatique, Klincksiek, Paris 1968).

/8/ : DELAVENAY : *La machine à traduire* (Mouton RCO 1964). (Collection : Que sais-je).

/9/ : G. MOUNIN : *La machine à traduire* (Mouton RCO 1964).

/10/ : *Computational linguistic at Rand* (1967).

Documentation technique : 1) Emanant de la C.I.I.

/11/ : *COBOL sous SIRIS 7* - manuel d'opérations (Réf : 3908 E/Fr).

/12/ : *COBOL 65 C.I.I. sous SIRIS 7* - manuel d'utilisation (Réf : 3760 E1/F

/13/ : *TRI-FUSION sous SIRIS 7* - manuel d'opérations et d'utilisation

2) Emanant de l'I.U.C.A.

/14/ : *Note de service n° 5*, Exploitation des travaux sous SIRIS 7.

/15/ : *Note technique n° 9.* Différences entre le COBOL sous BPM et le COBOL sous SIRIS 7 (Réf : EXPCOB 0/03/09).

/16/ : *Note technique n° 10.* Codes d'exception détectés par les services du moniteur SIRIS 7 (Réf : EXPSS7 0/04/10).

/17/ : *Note technique n° 17.* Utilisation du moniteur SIRIS 7. Langage de commande (Réf : EXPSS7 0/05/17).

/18/ : *Note technique n° 18.* Utilisation SIRIS 7. Notions et exemples d'utilisation (Réf : ZXPSS7 0/05/18).

-:-

NOM DE L'ETUDIANT : JOUIN Jean Léon

NATURE DE LA THESE : Doctorat de Spécialité en Mathématiques Appliquées.

VU, APPROUVE

& PERMIS D'IMPRIMER

NANCY, le 11 octobre 1973.

LE PRESIDENT DE L'UNIVERSITE DE NANCY I

J.R. HELLUY