HAL
open science

Une méthode d'indexation automatique des documents fondée sur l'exploitation de leurs propriétés structurelles : application à un corps technique

Dalila Kerkouba

▶ **To cite this version:**

Dalila Kerkouba. Une méthode d'indexation automatique des documents fondée sur l'exploitation de leurs propriétés structurelles : application à un corps technique. Modélisation et simulation. Institut National Polytechnique de Grenoble - INPG, 1984. Français. NNT : . tel-00312247

HAL Id: tel-00312247

https://theses.hal.science/tel-00312247v1

Submitted on 25 Aug 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THESE

présentée à

l'Institut National Polytechnique de Grenoble

pour obtenir le grade de

DOCTEUR DE TROISIEME CYCLE

Informatique

par

Dalila KERKOUBA

UNE METHODE D'INDEXATION AUTOMATIQUE DES DOCUMENTS FONDEE SUR L'EXPLOITATION DE LEURS PROPRIETES STRUCTURELLES.

APPLICATION A UN CORPUS TECHNIQUE.

Thèse soutenue le 22 novembre 1984 devant la Commission d'Examen :

Monsieur	J. MOSSIERE	:	Président
Messieurs	M. ADIBA E. ANDRE Y. CHIARAMELLA E. CHOURAQUI		Examinateurs

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

Année universitaire 1982-1983

Président de l'Université : D. BLOCH

Vice-Président : René CARRE
Hervé CHERADAME
Marcel IVANES

PROFESSEURS DES UNIVERSITES :

ANCEAU François	E.N.S.I.M.A.G.
BARRAUD Alain	E.N.S.I.E.G.
BAUDELET Bernard	E.N.S.I.E.G.
BESSON Jean	E.N.S.E.E.G.
BLIMAN Samuel	E.N.S.E.R.G.
BLOCH Daniel	E.N.S.I.E.G.
BOIS Philippe	E.N.S.H.G.
BONNETAIN Lucien	E.N.S.E.E.G.
BONNIER Etienne	E.N.S.E.E.G.
BOUVARD Maurice	E.N.S.H.G.
BRISSONNEAU Pierre	E.N.S.I.E.G.
BUYLE BODIN Maurice	E.N.S.E.R.G.
CAVAIGNAC Jean-François	E.N.S.I.E.G.
CHARTIER Germain	E.N.S.I.E.G.
CHENEVIER Pierre	E.N.S.E.R.G.
CHERADAME Hervé	U.E.R.M.C.P.P.
CHERUY Arlette	E.N.S.I.E.G.
CHIAVERINA Jean	U.E.R.M.C.P.P.
COHEN Joseph	E.N.S.E.R.G.
COUMES André	E.N.S.E.R.G.
DURAND Francis	E.N.S.E.E.G.
DURAND Jean-Louis	E.N.S.I.E.G.
FELICI Noël	E.N.S.I.E.G.
FOULARD Claude	E.N.S.I.E.G.
GENTIL Pierre	E.N.S.E.R.G.
GUERIN Bernard	E.N.S.E.R.G.
GUYOT Pierre	E.N.S.E.E.G.
IVANES Marcel	E.N.S.I.E.G.
JAUSSAUD Pierre	E.N.S.I.E.G.
JOUBERT Jean-Claude	E.N.S.I.E.G.
JOURDAIN Geneviève	E.N.S.I.E.G.
LACOUME Jean-Louis	E.N.S.I.E.G.
LATOMBE Jean-Claude	E.N.S.I.M.A.G.

.../...

LESSIEUR Marcel	E.N.S.H.G.
LESPINARD Georges	E.N.S.H.G.
LONGEQUEUE Jean-Pierre	E.N.S.I.E.G.
MAZARE Guy	E.N.S.I.M.A.G.
MOREAU René	E.N.S.H.G.
MORET Roger	E.N.S.I.E.G.
MOSSIERE Jacques	E.N.S.I.M.A.G.
PARIAUD Jean-Charles	E.N.S.E.E.G.
PAUTHÈNET René	E.N.S.I.E.G.
PERRET René	E.N.S.I.E.G.
PERRET Robert	E.N.S.I.E.G.
PIAU Jean-Michel	E.N.S.H.G.
POLOUJADOFF Michel	E.N.S.I.E.G.
POUPOT Christian	E.N.S.E.R.G.
RAMEAU Jean-Jacques	E.N.S.E.E.G.
RENAUD Maurice	U.E.R.M.C.P.P.
ROBERT André	U.E.R.M.C.P.P.
ROBERT François	E.N.S.I.M.A.G.
SABONNADIERE Jean-Claude	E.N.S.I.E.G.
SAUCIER Gabrielle	E.N.S.I.M.A.G.
SCHLENKER Claire	E.N.S.I.E.G.
SCHLENKER Michel	E.N.S.I.E.G.
SERMET Pierre	E.N.S.E.R.G.
SILVY Jacques	U.E.R.M.C.P.P.
SOHM Jean-Claude	E.N.S.E.E.G.
SOUQUET Jean-Louis	E.N.S.E.E.G.
VEILLON Gérard	E.N.S.I.M.A.G.
ZADWORNY François	E.N.S.E.R.G.

PROFESSEURS ASSOCIES

BASTIN Georges	E.N.S.H.G.
BERRIL John	E.N.S.H.G.
CARREAU Pierre	E.N.S.H.G.
GANDINI Alessandro	U.E.R.M.C.P.P.
HAYASHI Hirashi	E.N.S.I.E.G.

PROFESSEURS UNIVERSITE DES SCIENCES SOCIALES (Grenoble II)

BOLLIET Louis
Chatelin Françoise

PROFESSEURS E.N.S. Mines de Saint-Etienne

RIEU Jean
SOUSTELLE Michel

CHERCHEURS DU C.N.R.S.

FRUCHART Robert	Directeur de Recherche
VACHAUD Georges	Directeur de Recherche

.../...

ALLIBERT Michel	Maître de Recherche
ANSARA Ibrahim	Maître de Recherche
ARMAND Michel	Maître de Recherche
BINDER Gilbert	
CARRE René	Maître de Recherche
DAVID René	Maître de Recherche
DEPORTES Jacques	
DRIOLE Jean	Maître de Recherche
GIGNOUX Damien	
GIVORD Dominique	
GUELIN Pierre	
HOPFINGER Emil	Maître de Recherche
JOUD Jean-Charles	Maître de Recherche
KAMARINOS Georges	Maître de Recherche
KLEITZ Michel	Maître de Recherche
LANDAU Ioan-Dore	Maître de Recherche
LASJAUNIAS J.C.	
MERMET Jean	Maître de Recherche
MUNIER Jacques	Maître de Recherche
PIAU Monique	
PORTESEIL Jean-Louis	
THOLENCE Jean-Louis	
VERDILLON André	

CHERCHEURS du MINISTERE de la RECHERCHE et de la TECHNOLOGIRE (Directeurs et Maîtres de Recherches, ENS Mines de St. Etienne)

LESBATS Pierre	Directeur de Recherche
BISCONDI Michel	Maître de Recherche
KOBYLANSKI André	Maître de Recherche
LE COZE Jean	Maître de Recherche
LALAUZE René	Maître de Recherche
LANCELOT Francis	Maître de Recherche
THEVENOT François	Maître de Recherche
TRAN MINH Canh	Maître de Recherche

PERSONNALITES HABILITEES à DIRIGER des TRAVAUX de RECHERCHE (Décision du Conseil Scientifique)

ALLIBERT Colette	E.N.S.E.E.G.
BERNARD Claude	E.N.S.E.E.G.
BONNET Rolland	E.N.S.E.E.G.
CAILLET Marcel	E.N.S.E.E.G.
CHATILLON Catherine	E.N.S.E.E.G.
CHATILLON Christian	E.N.S.E.E.G.
COULON Michel	E.N.S.E.E.G.
DIARD Jean-Paul	E.N.S.E.E.G.
EUSTAPOPOULOS Nicolas	E.N.S.E.E.G.
FOSTER Panayotis	E.N.S.E.E.G.

.../...

GALERIE Alain	E.N.S.E.E.G.
HAMMOU Abdelkader	E.N.S.E.E.G.
MALMEJAC Yves	E.N.S.E.E.G. (CENG)
MARTIN GARIN Régina	E.N.S.E.E.G.
NGUYEN TRUONG Bernadette	E.N.S.E.E.G.
RAVAINE Denis	E.N.S.E.E.G.
SAINFORT	E.N.S.E.E.G. (CENG)
SARRAZIN Pierre	E.N.S.E.E.G.
SIMON Jean-Paul	E.N.S.E.E.G.
TOUZAIN Philippe	E.N.S.E.E.G.
URBAIN Georges	E.N.S.E.E.G. (Laboratoire des ultra-réfractaires ODEILLON)
GUILHOT Bernard	E.N.S. Mines Saint Etienne
THOMAS Gérard	E.N.S. Mines Saint Etienne
DRIVER Julien	E.N.S. Mines Saint Etienne
BARIBAUD Michel	E.N.S.E.R.G.
BOREL Joseph	E.N.S.E.R.G.
CHOVET Alain	E.N.S.E.R.G.
CHEHIKIAN Alain	E.N.S.E.R.G.
DOLMAZON Jean-Marc	E.N.S.E.R.G.
HERAULT Jeanny	E.N.S.E R.G.
MONLLOR Christian	E.N.S.E.R.G.
BORNARD Guy	E.N.S.I.E.G.
DESCHIZEAU Pierre	E.N.S.I.E.G.
GLANGEAUD François	E.N.S.I.E.G.
KOFMAN Walter	E.N.S.I.E.G.
LEJEUNE Gérard	E.N.S.I.E.G.
MAZUER Jean	E.N.S.I.E.G.
PERARD Jacques	E.N.S.I.E.G.
REINISCH Raymond	E.N.S.I.E.G.
ALEMANY Antoine	E.N.S.H.G.
BOIS Daniel	E.N.S.H.G.
DARVE Félix	E.N.S.H.G.
MICHEL Jean-Marie	E.N.S.H.G.
OBLED Charles	E.N.S.H.G.
ROWE Alain	E.N.S.H.G.
VAUCLIN Michel	E.N.S.H.G.
WACK Bernard	E.N.S.H.G.
BERT Didier	E.N.S.I.M.A.G.
CALMET Jacques	E.N.S.I.M.A.G.
COURTIN Jacques	E.N.S.I.M.A.G.
COURTOIS Bernard	E.N.S.I.M.A.G.
DELLA DORA Jean	E.N.S.I.M.A.G.
FONLUPT Jean	E.N.S.I.M.A.G.
SIFAKIS Joseph	E.N.S.I.M.A.G.
CHARUEL Robert	U.E.R.M.C.P.P.
CADET Jean	C.E.N.G.
COEURE Philippe	C.E.N.G. (LETI)

.../...

DELHAYE Jean-Marc	C.E.N.G. (STT)
DUPUY Michel	C.E.N.G. (LETI)
JOUVE Hubert	C.E.N.G. (LETI)
NICOLAU Yvan	C.E.N.G. (LETI)
NIFENECKER Hervé	C.E.N.G.
PERROUD Paul	C.E.N.G.
PEUZIN Jean-Claude	C.E.N.G. (LETI)
TAIEB Maurice	C.E.N.G.
VINCENDON Marc	C.E.N.G.

LABORATOIRES EXTERIEURS

DEMOULIN Eric	C.N.E.T.
DEVINE	C.N.E.T. (R.A.B.)
GERBER Roland	C.N.E.T.
MERCKEL Gérard	C.N.E.T.
PAULEAU Yves	C.N.E.T.
GAUBERT C.	I.N.S.A. Lyon

Je tiens à remercier ici:

Monsieur Jacques MOSSIERE, professeur à l'INPG, qui a accepté de présider le jury de cette thèse;

Monsieur Michel ADIBA, professeur à l'USMG, pour sa participation au jury;

Monsieur Yves CHIARAMELLA, professeur à l'USMG, sans qui ce document n'aurait jamais vu le jour. Qu'il trouve ici ma profonde reconnaissance pour m'avoir suivie tout au long de ce travail par des critiques constructives et des efforts constants, sans oublier la patience qu'il a eu pour lire et corriger à maintes reprises le manuscrit de cette thèse;

Monsieur Edouard ANDRE, directeur du projet CONCERTO, pour sa participation au jury et pour l'intérêt qu'il a porté à ce travail;

Monsieur Eugène CHOURAQUI, chargé de recherche au CNRS à MARSEILLE, qui m'a fait l'honneur de participer au jury de thèse;

Madame Marie-Françoise BRUANDET, maitre assistant à l'USSG, pour l'aide qu'elle m'a apportée tout au long de ce travail et pour l'intérêt qu'elle y a accordé.

Le travail présenté dans cette thèse est le fruit de la participation et de la bonne entente de tous les membres de l'équipe. Qu'ils trouvent ici mes sincères remerciements.

Enfin je tiens à remercier profondément Bruno DEFUDE, Orieta SANTANA et Françoise RENZETTI, les deux premiers pour l'aide précieuse dans la mise en forme de ce manuscrit, la dernière pour le mal qu'elle s'est donnée à indexer manuellement certains textes utilisés dans l'expérimentation de ce travail.

Ce travail a été financé par le CNET Lannion dans le cadre du projet CONCERTO.

TABLE DES MATIERES

-ooo-

I - INTRODUCTION 5

1. PROBLEME DE L'INDEXATION 5

1.1. Fonctionnalités d'un système de recherche d'information (SRI) 5

1.2. Définition de l'indexation 8

1.2.1. Contenu d'un corpus documentaire 9

1.2.2. Caractéristiques d'une bonne recherche documentaire 11

1.2.3. Conclusion 12

2. EVALUATION D'UN SRI 14

3. APPORTS PROPOSES 15

4. LES REALISATIONS 18

II- STRUCTURATION DE LA CONNAISSANCE DANS UN TEXTE 21

1. HIERARCHIE DES CONSTITUANTS TEXTUELS 22

1.1. Choix d'organisation de la connaissance dans notre application 23

1.2. Redéfinition du corpus documentaire 24

2. INDICATEURS PARTICULIERS DANS UN DOCUMENT 26

2.1. Titre d'une entité 26

2.1.1. Titres-procédés 27

2.1.2. Titres informatifs 28

2.2. Typographie particulière de certains termes du texte 28

3. LE TEXTE EN LANGUE NATURELLE 29
3.1. Liaison Syntaxe-Sémantique 29
3.2. Définition informelle des GCP 33

4. CONCLUSION 35

III-METHODES D'INDEXATION EXISTANTES 37

1. PREAMBULE 37
1.1. Définition et rôle du thésaurus à l'indexation 38
1.2. Problème de précision et de rappel 40

2. METHODES STATISTIQUES ET METHODES LINGUISTIQUES 41
2.1. Méthodes Statistiques 41
2.2. Insuffisances des Méthodes Statistiques 44
2.3. Méthodes Linguistiques 47

3. MISE EN OEUVRE DANS TROIS SYSTEMES OPERATIONNELS TYPES 51
3.1. Aspects Linguistiques 53
3.1.1. Analyse Morpho-Syntaxique 53
3.1.2. Analyse Sémantique 55
3.2. Aspects Statistiques 57
3.3. Aspects Sémantiques 60
3.4. Conclusion 62

IV- STRATEGIE D'UTILISATION DES OUTILS LINGUISTIQUES 65

1. ANALYSE MORPHO-SYNTAXIQUE 65
1.1. Formes d'un GCP 66
1.2. Outils linguistiques envisagés 69

2. ROLE DU THESAURUS 72
2.1. Structure actuelle du thésaurus dans l'application 72

2.2. Trace pseudo-syntaxique d'une clique 76
2.3. Pattern-matching entre un GCP et le thésaurus 77

3. CASSURE SYNTAXIQUE D'UN GCP 79
3.1. Stratégie de cassure d'un GCP 80
3.2. Le réseau syntaxique statique 82
3.2.1. Définition 82
3.2.2. Utilisation du réseau statique 83
3.3. Le réseau dynamique 84

4. INDEXATION IMPLICITE 86

V - CRITERES STATISTIQUES 89

1. PRINCIPES GENERAUX 89

2. PONDERATION DE LA RELATION D'INDEXATION 91
2.1. Définitions préalables 91
2.2. Représentativité mutuelle entre un terme et une unité d'indexation 92
2.2.1. Représentativité de d par rapport à t 92
2.2.2. Représentativité de t par rapport à d 94
2.2.3. Représentativité mutuelle entre t et d 95

VI- INDEXATION DYNAMIQUE 97

1. PRESENTATION 97
2. ANALYSE D'UNE UNITE D'INDEXATION MINIMALE 98
3. STRATEGIE DE REMONTEE DES TERMES D'INDEXATION 99
4. VALUATION DU POIDS D'INDEXATION DANS LA STRATEGIE DE REMONTEE 100
5. CONCLUSION 104

VII-REALISATIONS 107

1. INTRODUCTION 107

2. DESCRIPTION ET STOCKAGE DU TEXTE PAR STRUCTDOC 108

3. OUTILS LINGUISTIQUES 110
3.1. Lemmatisation du vocabulaire 110
3.2. L'analyse syntaxique de surface via ANAL 111

4. LE MODULE D'INDEXATION 114

5. CONCLUSION 117

VIII-EXPERIMENTATION 121

1. LE CORPUS TRAITE 121

2. LA BASE DE CONNAISSANCES 122

3. RESULTATS DE L'INDEXATION 125
3.1. Evaluation quantitative de la méthode 125
3.2. Evaluation qualitative de la méthode 129
3.3. Indexation dynamique 144
3.4. Conclusion 147

IX- CONCLUSION 149

BIBLIOGRAPHIE 153

ANNEXE:
ALGORITHMES DES PRINCIPALES FONCTIONS DU MODULE D'INDEXATION

CHAPITRE I

INTRODUCTION

1. PROBLEME DE L'INDEXATION

1.1. Fonctionnalités d'un système de recherche d'information (SRI)

Un système documentaire est composé d'un ensemble de fonctions qui permettent de gérer et de manipuler des documents regroupés dans une base.
Cet ensemble de documents constitue un corpus qui traite d'un thème particulier ou d'un ensemble de thèmes donnés et intéresse un ensemble d'usagers.
Le rôle d'un système documentaire [RIJ 79], [SAL 75], [SAL 83] est de permettre aux usagers un accès au corpus via des requêtes concernant tel ou tel thème particulier; l'objectif visé se résume en l'établissement d'une communication entre la base de documents et les utilisateurs.
Etant donné un usager recherchant une information partielle relative à un thème précis, il exprime ce thème dans une requête d'interrogation et le système repère et procure le ou les documents susceptibles de satisfaire son besoin (i.e considérés comme caractéristiques de ce thème).
Dans le cas où l'information obtenue est satisfaisante, ce sous ensemble de documents est considéré comme pertinent.
On peut distinguer, suivant les systèmes, deux formes possibles de réponses à une requête documentaire:

- La réponse est composée d'un ensemble de références vers les documents sélectionnés. Ce type de réponse est fourni par les systèmes où la base documentaire est constituée uniquement de références à l'ensemble des documents [MIS 78], [GOL 72]. De

tels systèmes correspondent le plus souvent à des bases de données factuelles classiques.

- La réponse est composée de l'ensemble des textes des documents pertinents; ce type de réponse ne peut être obtenu que dans le cas où l'on dispose effectivement des textes complets (base de données textuelles).
  Ceci se pratique lorsque la recherche se fait directement sur les textes des documents [STA 79], ou dans le cas où la manipulation des documents ne concerne pas uniquement la recherche documentaire mais un ensemble d'applications relevant d'un domaine précis, [BCH 81], [HAM 81], [JOL 81], [LAN 81], [TIG 82]. (EX: bureautique, bases de données généralisées...).
  Dans ces domaines la documentation est vue comme une fonction particulière parmi d'autres.

En général, lorsqu'une base recouvre plusieurs thèmes, elle est subdivisée en autant de sous bases qui contiennent chacune l'ensemble des documents du corpus relatifs à un thème particulier. Dans ce cas, une requête d'usager peut être de deux types [MIS 78]:

- La requête porte sur un thème bien précis (spécifié par l'usager), elle concerne seulement la sous base du système relative au thème en question.
  La réponse est extraite uniquement de cette sous base.

- La requête concerne toute la base documentaire (pas de spécification de thème a priori); la consultation de chaque sous base est nécessaire et indépendante des autres.
  La réponse, dans ce cas, est constituée de l'ensemble des réponses, extraites de chaque sous base. Tout se déroule donc comme s'il s'agissait de consultations de bases documentaires différentes.

Dans les deux cas, le problème est donc ramené à la consultation d'une base relative à un thème particulier; c'est la raison pour laquelle on ne s'intéresse qu'à cette hypothèse dans l'étude qui suit. L'ensemble des documents de la base forme alors un corpus homogène.

Pour sélectionner un document de la base susceptible de satisfaire une requête, le système devrait, en principe, procéder en deux étapes:

- prendre connaissance du contenu du document;
- examiner si ce contenu correspond à la requête.

Il est évident qu'il est impossible de traiter ainsi les requêtes sur tous les textes (temps et coût insupportables), d'où la nécessité d'introduire une étape préalable d'analyse des documents, ou indexation.

Le résultat de l'indexation est une représentation des documents, sous une forme réduite mais fidèle qui permettra de les désigner lors de l'interrogation de la base [AND 73A], [BOO 75], [COO 78], [MAR 77], [RIJ 79], [SAL 71], [SPA 71], [WAL 79].

De toutes ces considérations, il ressort qu'un système documentaire peut être caractérisé par deux fonctions principales qui sont étroitement liées: la fonction ACQUISITION et la fonction INTERROGATION.

Fonction ACQUISITION

C'est l'ensemble des opérations que subissent les documents avant leur mémorisation dans la base documentaire .
L'opération la plus importante est l'indexation documentaire: elle vise à reconnaître et à représenter le contenu

caractéristique d'un document, sous une forme adaptée à l'utilisation qui en sera faite ultérieurement lors de l'interrogation.
L'automatisation de cette opération constitue le sujet de notre travail; les méthodes proposées ainsi que les outils informatiques correspondants que nous avons réalisés et expérimentés, sont présentés dans les chapitres qui suivent.

Fonction INTERROGATION

Elle correspond, comme indiqué plus haut, à la finalité essentielle du logiciel documentaire, et comprend toutes les possibilités que le logiciel offre pour satisfaire l'utilisateur face à ses problèmes de recherche d'informations.
Le coeur de cette fonction est donc l'opération de recherche documentaire : elle consiste à repérer, à partir de certains critères, un ensemble de documents pertinents parmi l'ensemble de la base .
Ces critères, dits de recherche, permettent de désigner d'une manière directe ou indirecte (cf III), la représentation interne du document. De ce fait le repérage correct du document n'est possible que si ces critères de recherche ont été pris en compte lors de l'indexation du document. La qualité de cette fonction est donc étroitement liée à la manière dont a été analysé et indexé le document.

1.2. Définition de l'indexation

De cette brève présentation des fonctionnalités d'un SRI, il ressort que le rôle de l'indexation peut se résumer ainsi:
l'indexation permet de représenter le contenu du document dans une expression sémantiquement équivalente mais restreinte au sous ensemble des notions les plus caractéristiques du document, dans le but d'aiguiller le processus d'interrogation uniquement

vers les documents les plus pertinents relativement à l'information recherchée par l'usager.
Ce formalisme est donc lié aux modalités d'interprétation de la requête pour l'interrogation de la base: répondre à une requête, revient à réaliser une sorte de "pattern-matching" entre l'expression réduite des documents et le contenu sémantique de la requête.

Le pattern-matching réussit si l'expression de la requête est comprise ou peut être déduite de l'expression du document ; autrement dit pour que cette opération soit possible, il faut que dans l'expression du contenu du document on ait tenu compte des expressions possibles d'une requête.

Ceci revient en fait à établir une relation a priori entre le contenu du corpus documentaire et le contenu des requêtes pouvant être émises par les usagers. Dans la définition de cette relation d'indexation, il est crucial que les propriétés du langage d'interrogation ne soient pas ignorées.

Avant de définir cette relation, on présentera quelques propriétés liées au contenu d'un corpus ; on précisera leur utilité vis à vis d'un système documentaire. En particulier, on essaiera de dégager les objectifs à atteindre pour une bonne recherche documentaire et d'en déduire les caractéristiques de l'indexation.

1.2.1. Contenu d'un corpus documentaire

Un corpus documentaire homogène est donc un ensemble de documents qui traitent d'un thème particulier commun.
Un thème est décrit par un ensemble de concepts, chacun d'entre eux étant considéré comme une notion particulière qui désigne partiellement ce thème.
En fait, le thème est souvent désigné d'une manière hiérarchique

(générique, spécifique) par cet ensemble de concepts.

La description du thème à travers les documents se fait par l'intermédiaire de cet ensemble de concepts. Chaque document concourt partiellement à la définition du thème, et porte donc sur un sous-ensemble de ces concepts. L'information partielle véhiculée par le document sur le thème, s'exprime par ces concepts.

En particulier on distingue:

- Les concepts génériques
  Un concept générique est souvent considéré comme un sous thème du thème traité; de ce fait la description du thème dans un document peut être assimilée à celle des concepts génériques qu'il contient.
  Un concept générique étant une notion large peut être à son tour décrit à travers les documents par un sous-ensemble de notions spécifiques plus précises.

  EX: dans un corpus médical, le concept "maladie" est un concept générique qui peut introduire des concepts plus spécifiques tels que "maladie du coeur", "maladie du système nerveux"...

  Un concept générique est généralement caractérisé par sa forte évocation à travers les documents du corpus; il peut uniquement servir à rappeler le thème traité, ou, plus exactement, à rattacher l'information apportée par un texte au thème principal.

- Les concepts spécifiques
  Leur utilisation est plus précise et beaucoup plus restreinte; ils sont donc intéressants par les raffinements qu'ils apportent dans la définition du thème principal et des

concepts génériques.
Dans un document, les concepts spécifiques représentent l'information originale (par rapport au reste du corpus), apportée sur la définition du thème.

On peut donc résumer ainsi la relation entre les documents et l'ensemble des concepts décrivant le thème:

- L'information présente dans chaque document porte, avec des degrés différents, sur un sous-ensemble de concepts du thème. On peut en déduire une notion de représentativité d'un document par rapport aux éléments du thème principal.

- Chaque concept du thème concourt, avec des degrés différents, dans l'expression et l'originalité de l'information véhiculée par un document du corpus.
  On peut en déduire une notion de représentativité d'un concept par rapport au contenu de chaque document.

1.2.2. Caractéristiques d'une bonne recherche documentaire

Un usager intéressé par un aspect particulier d'un thème, formule sa requête à l'aide d'un sous ensemble de concepts de ce thème qui lui paraissent les mieux indiqués pour désigner l'information recherchée.
L'information obtenue en réponse est estimée pertinente si l'usager juge qu'elle traite des aspects correspondants à son besoin précis, et si elle est la plus complète possible (sans pour autant dépasser en volume ses capacités individuelles d'appréhension).

A ce stade, il faut remarquer que les besoins des usagers sont très nuancés; en particulier on peut distinguer deux profils d'usagers: l'usager qui s'initie au thème (le non spécialiste), et l'usager qui possède déjà des connaissances approfondies sur

le thème (le spécialiste).

Intéressés par le même aspect d'un thème, l'information recherchée par l'un peut être différente de l'information recherchée par l'autre.
L'usager non spécialiste, étant novice, souhaiterait dans la majorité des cas avoir une vue générale sur le thème, alors que l'usager spécialiste souhaiterait au contraire recevoir une information plus détaillée et plus précise.

Il est donc évident qu'il faut chercher à adapter le degré de précision des réponses à leur compétence dans le thème exploré.

1.2.3. Conclusion

La mise en correspondance effective entre requêtes et documents doit donc être établie via l'ensemble des concepts décrivant le thème en tenant compte du fait que:

- Au niveau de l'expression de la requête, les notions sont choisies selon un besoin d'information particulier.

- Au niveau du document, concepts et contenu effectif (texte) sont liés selon un degré de représentativité mutuelle.

Dans la relation entre concepts et documents, on ne peut ignorer les deux problèmes critiques auxquels on est confronté en recherche documentaire: la représentativité et le pouvoir discriminant des concepts par rapport au corpus, [SAL 75], [SAL 83].

- Représentativité d'un concept par rapport à un document: on évalue ici dans quelle mesure le concept participe à l'information véhiculée par le document.

Un concept peut exister dans un document sans pour autant être caractéristique de son contenu. Ce document ne serait pas considéré comme pertinent s'il était fourni en réponse à une question axée autour de ce concept.

- Discriminance d'un document par un concept: ceci évalue dans quelle mesure le concept participe à l'originalité du document par rapport au reste du corpus ou, autrement dit, dans quelle mesure le document décrit complétement le concept.

Pour être efficace à la recherche, la discriminance du concept par rapport au corpus doit donc aller dans le même sens que sa représentativité par rapport aux documents. En d'autres termes, un concept efficace pour la recherche est fortement représentatif d'un très petit nombre de documents.
En effet, un concept qui n'apparaît qu'une seule fois dans le corpus, discrimine totalement le document qui le contient, mais il peut être inexploitable à la recherche s'il n'est pas représentatif de ce document.
La discriminance d'un concept est d'autant plus forte que l'utilisation de ce concept dans le corpus se limite à un petit nombre de documents; en fait, la discriminance peut donc être rattachée à la notion de représentativité du document par rapport au concept.

Pour les besoins de l'interrogation, la relation à établir entre un concept du thème et un document du corpus doit donc tenir compte de la représentativité du concept par rapport au document, et de la représentativité du document par rapport au concept.
En définitive, nous utiliserons les définitions suivantes de la notion de représentativité mutuelle document_concept:

- La représentativité du document par rapport au concept évalue le degré de description du concept dans ce document, compte tenu de sa description dans la totalité du corpus.

- La représentativité du concept par rapport au document évalue le degré de participation de ce concept dans l'information véhiculée par le document, compte tenu de l'ensemble des concepts évoqués dans le document.

Répondre de manière satisfaisante à une requête, c'est exploiter cette représentativité mutuelle entre concept et document selon le besoin émis.

2. EVALUATION D'UN SRI

En pratique, mise à part la rapidité, l'efficacité d'un système est toujours mesurée par le taux de bruit et de silence par rapport à des requêtes documentaires précises.

Le bruit est le taux de documents non pertinents sélectionnés dans la réponse; il évalue donc le degré de non pertinence de la réponse.
La pertinence (relevance) [SAL 75], [SAL 83], indique la façon dont le système répond au besoin réel en information de l'usager et de ce fait s'exprime essentiellement par le degré de satisfaction des utilisateurs. Elle est fonction de divers facteurs: qualité et cohérence de l'indexation des documents, indexation et formulation de la requête, différence dans l'ampleur d'une même notion considérée dans le document et dans la requête...

Le silence est le taux de documents pertinents à une question, non retrouvés lors de l'interrogation de la base; il évalue donc le degré de non exhaustivité de la réponse .

Pour remédier au silence, on effectue généralement une reformulation de la requête: on réinterroge la base en élargissant la requête.
Cette opération (relevance feedback), [SAL 75], [SAL 83], est effectuée lorsque l'usager estime que la réponse à une première requête est insuffisante; elle permet d'extraire de la base davantage de documents se rapportant à un thème.
Pour l'élargissement de la requête, on utilise généralement les environnements sémantiques (synonymes, génériques, spécifiques, etc...) des concepts primaires de la requête (cf III)
A ce niveau, l'évaluation d'un système est également mesurée par le taux de rappel (recall), [SAL 75], [SAL 83], par rapport à des requêtes d'usagers: c'est le taux de documents pertinents retrouvés pour la requête par rapport à l'ensemble des documents pertinents de la base.

Un bon système documentaire est un système qui fournit un taux de bruit et un taux de silence aussi faibles que possible, mais souvent ces deux objectifs sont contradictoires: en tentant de réduire l'un, on peut augmenter l'autre.

3. APPORTS PROPOSES

Nous résumons ci-dessous l'ensemble des points méthodologiques essentiels sur lesquels est fondée la stratégie d'indexation automatique proposée, et qui constituent notre contribution à la résolution de ce problème.

1) NOTION DE TERME D'INDEXATION

En analyse documentaire, le fond du problème reste de nature sémantique.
De nombreux travaux dans ce domaine [DEB 82], [DEW 81], [FLU 77], [GUE 82], ont porté sur le fait que SYNTAXE et

SEMANTIQUE sont deux disciplines étroitement liées, et il devient évident que l'étude des liens syntaxiques entre les mots d'une phrase donnée constitue une première approche quant à la production du sens véhiculé par cette phrase.

Dans plusieurs SRI existants [DEB 82], [DEW 81], [FLU 77],[SYN 74], on ne s'intéresse plus aux mots clés isolés classiques mais à des syntagmes beaucoup plus riches du point de vue sémantique, pour indexer les textes en langue naturelle.
Parmi les divers syntagmes du texte, l'intérêt est porté sur un sous ensemble des syntagmes nominaux, qui à notre avis, véhiculent une information suffisante dans un cadre documentaire.
Cette approche nécessite un renforcement de certains aspects linguistiques qui sont développés en II-2.

2) PRISE EN COMPTE DE LA STRUCTURE GLOBALE DU TEXTE

L'information enregistrée dans un texte peut parfois constituer un 'amalgame' d'objets de natures différentes.
Par exemple, dans une notice technique, on peut trouver du texte en langue naturelle, des programmes écrits dans des langages de programmation différents, des formules mathématiques, des schémas...
Au niveau externe, cet ensemble d'objets est présenté selon des conventions de visualisation particulières, qui peuvent être décrites chacune par un formalisme propre [JOL 81], [TIG 82], [TIG 83], [VIR 81], [VIR 82b].
L'ensemble de ces composants est présenté selon une structure hiérarchique classique en chapitres, sous-chapitres, paragraphes, sous-paragraphes.
Cette structure a été déterminée par l'auteur: elle dénote ses intentions dans l'organisation des connaissances exposées dans le texte, et ne peut être changée sans altérer la

compréhension du contenu du document. En particulier, certains composants dans cette hiérarchie (titres, introduction, conclusion, ...) jouent un rôle spécifique pour la présentation de l'ensemble [JOL 81], [TIG 82], [TIG 83], [VIR 81], [VIR 82B].
Dans la stratégie proposée, on s'intéresse à cet aspect structurel du texte pour l'indexation automatique; ceci est exposé en II.1.

3) REFERENCE A UN THESAURUS

Pour l'indexation des textes, on exploite également un thésaurus, construit automatiquement à partir des documents, ([BRU 80A], [BRU 81], [BRU 82]) ; son contenu constitue une base de connaissance limitée au corpus traité.
Le thésaurus est utilisé à deux niveaux dans le processus d'indexation: pour le stockage des termes d'indexation d'une part, et pour un traitement linguistique assez spécifique (normalisation des termes via un pattern-matching) sur certains termes d'indexation d'autre part.
Ces aspects sont détaillés en III.3.

4) CRITERES STATISTIQUES POUR LA VALUATION DE LA RELATION D'INDEXATION

Des critères statistiques basés sur la fréquence d'occurrence des termes dans les éléments de texte permettent de définir, dans la méthode d'indexation proposée, une valuation de la relation d'indexation.
Plutôt qu'une sélection classique fondée sur des seuils prédéfinis et statistiques, nous proposons une fonction continue dans l'intervalle [0,1].
Cette solution offre plus de souplesse et permet en particulier de prendre en compte de manière efficace la mise à jour de l'indexation lorsque le corpus évolue.

Cette fonction est définie en V.4.

5) STRATEGIE D'INDEXATION DYNAMIQUE

Dans la méthode proposée, un corpus documentaire n'a pas la définition classique d'un ensemble de documents. L'unité textuelle sur laquelle portent l'indexation et l'interrogation est liée à la structure logique du document. Ceci permet de localiser plus précisément l'entité textuelle correspondant à la portée d'un concept, et d'adapter la réponse à une requête suivant un besoin précis: la réponse est un sous arbre de la structure globale du texte dont le niveau correspond au degré de précision de cette requête.
L'aspect dynamique permet également une réévaluation aisée de la relation d'indexation lorsque le corpus évolue. Ceci fait l'objet du chapitre VI.

4. LES REALISATIONS

Les réalisations ont porté sur le développement de trois outils principaux qui, à partir d'un texte initial, réalisent son analyse complète: Il s'agit des trois outils suivants:

1) L'outil de description de documents sous forme arborescente.

2) L'outil de traitement linguistique qui permet d'extraire d'un texte des éléments ayant la structure syntaxique retenue pour les termes d'indexation.

3) L'outil d'indexation propement dit, qui permet d'établir et de valuer la relation d'indexation entre un élément d'indexation et l'unité textuelle qui le contient.

Les programmes correspondants ont un caractère expérimental dont l'intérêt immédiat est la validation de la stratégie d'indexation proposée. Leur description est donnée en VII.

L'expérimentation de la méthode a porté sur un corpus technique d'environ 100.000 mots de texte en langue naturelle. Cette expérimentation s'insère dans le cadre du projet CONCERTO développé par le CNET-LANNION [AND 82A], [AND 82B]. Les premiers résultats obtenus sont exposés en VIII.

CHAPITRE II

STRUCTURATION DE LA CONNAISSANCE DANS UN TEXTE

Un document peut être composé d'un ensemble d'objets de natures différentes correspondant chacune à un formalisme de description particulier, [JOL 81], [LAN 83], [TIG 82], [TIG 83]. De nombreux types d'objets peuvent être décrits au moyen d'un outil syntaxique; citons notamment:

- type texte.
- type schéma.
- type graphique.
- type programme.
- type tableau.
  etc...

Le formalisme de chaque objet concourt à sa compréhension particulière et de la même manière, la structure de l'ensemble participe à la compréhension globale du document.

Dans la méthode d'indexation proposée, on s'intéresse à l'exploitation de la "méta-connaissance" qu'on peut extraire de certaines propriétés structurelles des textes:

- l'aspect syntaxique du texte en langue naturelle permet d'extraire les concepts "primaires" de son contenu.

- la structure globale du document permet d'extraire certains éléments utiles pour son indexation et de définir une stratégie d'indexation dynamique.

Dans ce domaine, et du point de vue informatique, de nombreux outils ont été développés pour la manipulation des divers formalismes d'un document, mais d'une manière indépendante:

- Analyseur syntaxique du langage naturel (cf III);
- Constructeurs et éditeurs de programmes [MEL 80], [ADA 81].
- Outils de manipulation de schémas et figures [KAH 81];
- Outils de manipulation de formules mathématiques [QUI 82]; etc...

La définition et l'utilisation des outils de manipulation de documents ont considérablement progressé ces dernières années, notamment dans les trois domaines de la bureautique, des bases de données généralisées et du traitement de texte [JOL 81], [MEL 83], [TIG 82], [TIG 83].
Dans ces applications informatiques, l'intérêt porte sur les diverses manipulations possibles du document dans un environnement bien délimité et par rapport à certaines fonctions prédéfinies. De ce fait, on distingue l'aspect conceptuel du document de son aspect purement physique.

Les recherches actuelles dans ce domaine portent sur deux points [JOL 81], [VIR 81], [VIR 82B], [LAN 81], [LAN 83], [MEL 83], [TIG 83].

- sur la définition d'un modèle de description du document, description qui unifie les différents aspects qui concernent a priori chacune des fonctions de manipulation.

- sur l'élaboration d'outils informatiques supportant un langage puissant pour la mise en oeuvre de cette description, et permettant l'intégration d'outils spécifiques à chaque formalisme.

1. HIERARCHIE DES CONSTITUANTS TEXTUELS

La structure logique du document est souvent liée implicitement à sa fonction et à son mode d'utilisation dans un environnement

bien défini: un contrat est souvent structuré de la même façon, et différemment d'un rapport de recherche, par exemple [JOL 81], [VIR 81], [TIG 82].
En général, on peut donc définir une structure logique figée d'information commune à un ensemble de documents manipulés dans un environnement précis; ceci permet de distinguer des CLASSES de documents: LETTRE, CONTRAT, RAPPORT...
Un document est défini par un type, qui le rattache à sa classe d'origine; à chaque classe de documents est donc également associé un type qui décrit la structure d'information commune à ces documents [JOL 81], [TIG 82], [TIG 83].
Un type de document porte un nom, relatif à la classe de documents associée (LETTRE, CONTRAT...).
Un type de document est défini par une grammaire qui donne l'ensemble de règles de composition de la structure hiérarchique du document.

1.1. Choix d'organisation de la connaissance dans notre application

Dans la stratégie proposée, la structure logique du document est considérée pour:

- redéfinir la notion de corpus documentaire selon des besoins précis dans une application.

- exploiter des indicateurs particuliers tels que les titres ou d'autres éléments porteurs d'informations utiles pour l'indexation du document.

Nous nous sommes intéressés pour l'instant au traitement d'une classe unique de documents du type 'notice technique', sur laquelle nous avons mené nos expérimentations (cf VIII).
Ce type est défini par 4 niveaux hiérarchiques en plus du niveau document (niveau 0). Les feuilles, ou éléments terminaux,

correspondent au contenu textuel du document et n'appartiennent plus au formalisme de la structure.
Par la suite, nous appellerons 'entité de structure' tout élément du document correspondant à un sous arbre de la structure logique.

La grammaire qui décrit la hiérarchie du type de documents du corpus traité peut être la suivante:

DOC ::= TITINF CHAP
CHAP ::= TITRE S-CHAP
S-CHAP::= TITRE PARAP
PARAP::= TITRE S-PARAP
S-PARAP::= TITRE SECTION
SECTION::= TITRE OBJETDOC
OBJETDOC::= TEXTE, PROGRAMME, SCHEMA, TABLEAU...
TITRE::= TITINF, TITSPEC
TITSPEC::= 'introduction','conclusion','généralité','résumé'
TITINF::= 'chaine de caractères'.

Nous ne nous intéressons qu'au traitement des objets de type TEXTE dans cette étude.

1.2. Redéfinition du corpus documentaire

Dans un SRI classique, l'entité textuelle élémentaire manipulée est le document: l'indexation porte sur un document et la réponse à une requête est la référence vers un document.
Pour des applications spécifiques et pour certains types de documents, ce genre de réponse s'avère peu précis, et la référence vers une entité textuelle moins dense du document serait souhaitable.
De manière symétrique, l'indexation devrait donc porter sur ces mêmes entités textuelles.
Dans notre approche, le corpus documentaire est redéfini en

fonction de la structure logique du document: c'est l'ensemble des entités de structure des documents analysés.
L'unité textuelle sur laquelle porte l'indexation et l'interrogation peut être située à n'importe quel niveau dans la structure logique du document.
Elle présente donc un aspect dynamique qui permet:

- de cerner le plus possible l'unité textuelle sur laquelle porte effectivement un concept, ce qui favorise la précision;

- un élargissement progressif (lors de l'interrogation de la base) de cette unité textuelle à des niveaux pères dans la structure du document, ce qui favorise un rappel adapté à des besoins particuliers.

L'unité textuelle ainsi définie (selon le type de document considéré) est appelée UNITE D'INDEXATION DYNAMIQUE.

L'aspect dynamique de cette unité est fonction du type de documents considéré; en particulier, on distingue deux niveaux de structure qui délimitent l'étendue matérielle possible d'une unité d'indexation:

- un niveau qui définit une arborescence minimale ou unité d'indexation minimale.
- un niveau qui définit une arborescence maximale ou unité d'indexation maximale.

Ces deux niveaux sont choisis a priori pour chaque type de document en fonction des besoins d'une application, et concerne en fait les niveaux limites de précision désirée lors de la recherche.

A titre d'exemple, nous avons fixé dans notre expérimentation l'unité maximale au niveau chapitre (niveau 1), car dans un

document du corpus traité les chapitres sont souvent indépendants. L'unité d'indexation minimale correspond au niveau sous-paragraphe (niveau 4) qui introduit les concepts spécifiques.

2. INDICATEURS PARTICULIERS DANS UN DOCUMENT

Ces indicateurs sont d'une part les titres, qui introduisent les entités de structure, d'autre part la typographie particulière de certains termes du texte.

2.1. Titre d'une entité

Un titre est choisi par l'auteur pour désigner le sujet traité dans le texte qui suit et est donc souvent informatif.
A cet effet R.FONDIN écrit: "le titre est un élément documentaire de premier ordre, puisque il est le reflet, à la fois du document et de son auteur; il apparaît la version la plus concise du contenu d'un document".
Mais il ajoute également que les résultats obtenus, en analysant uniquement les titres pour l'indexation de textes, n'ont pas toujours été satisfaisants et sont souvent fonction de la taille et de la nature du corpus traité.
Il remarque que dans un domaine scientifique le titre est moins fantaisiste que dans un domaine littéraire mais que même dans ce cas, quoi qu'on en dise, le titre contient toujours une information, même minime, sur le contenu du texte qui le suit [FON 82].

Dans notre étude, nous distinguons deux catégories de titres:

- les titres qui renseignent sur le type ou le rôle du texte qui suit, par rapport à la structure globale du document (titres-procédés) [VIR 81], [VIR 82b].

- les titres qui renseignent directement sur le contenu du texte qui suit (titres informatifs).

2.1.1. Titres-procédés

C'est le cas des titres tels que "INTRODUCTION", "RESUME", "CONCLUSION"...

L'introduction, par exemple, annonce brièvement ce qui va être développé dans le texte qui suit; bien que le titre "INTRODUCTION" n'apporte pas d'information précise en lui même, il indique que l'entité textuelle correspondante a un aspect résumant, par rapport à l'ensemble, d'où la nécéssité d'analyser d'une manière particulièrement fine cette entité dans laquelle des concepts essentiels seront présentés dans un contexte succint.

Un titre peut également renseigner sur le formalisme particulier d'une entité de structure: c'est le cas par exemple, de titres comme "PROGRAMME", "SCHEMA" dans un document technique, qui indiquent entre autres, un changement de formalisme dans le texte.

Les entités de structure introduites par ce type de titre lors de l'analyse des textes, doivent donc subir un traitement particulier et ne peuvent être indexées selon la même stratégie que le reste des entités du texte.

A cet effet, on définit une liste de titres de cette catégorie, appelée liste des titres-procédés; à chaque élément de cette liste, on associe un indicateur qui renseigne sur le traitement particulier que doit subir l'entité de structure qui le suit.

La liste des titres-procédés est une liste ouverte dépendant du corpus considéré.

Pour les textes techniques que nous traitons, elle contient les termes introductifs suivant:

LTP = [INTRODUCTION, GENERALITE, RESUME, CONCLUSION, PROGRAMME, SCHEMA, TABLEAU]

INTRODUCTION, GENERALITE, RESUME introduisent brièvement des concepts nouveaux qui sont décrits en détails dans les entités de structure qui suivent.

Dans une conclusion, qui a également un aspect résumant, il est souvent question de perspectives et donc de références vers des concepts d'autres entités de structures du corpus.

Les traitements particuliers envisagés pour chacun de ces titres seront vus en IV; ils sont partie intégrale de la stratégie de prise en compte de la structure du document pour l'indexation.

2.1.2. Titres informatifs

Un titre informatif est tout autre titre n'appartenant pas à la liste des titres-procédés; il renseigne directement sur le sujet décrit dans l'entité qui suit.

Dans notre méthode, tout titre de cette catégorie subit un traitement linguistique qui permet d'en extraire une forme syntagmatique principale. Cette forme (cf II-2) est systématiquement retenue comme terme d'indexation pour l'entité de structure qui est introduite par le titre.

2.2. Typographie particulière de certains termes du texte

Dans un texte, certains artifices purement graphiques dénotent une intention bien particulière de l'auteur: le fait de souligner ou d'encadrer un terme du texte, est souvent un signe

de mise en évidence de l'importance de ce terme dans le texte [JOL 81], [VIR 81].
Ces propriétés typographiques peuvent donc être exploitées pour l'indexation automatique du texte, mais doivent être utilisées avec précaution.
En effet, ils peuvent être source d'ambiguité du fait de leurs divers emplois dans les documents. Par exemple, l'interprétation du symbole '"' peut être différente d'un document à un autre; dans l'un, elle peut être effectivement relative au phénomène de mise en évidence, dans l'autre, elle peut introduire un terme en langue étrangère...

Dans un corpus où il n'existe pas une normalisation de cette typographie, il est impossible d'exploiter les propriétés qui en découlent.

Dans notre approche, on prévoit la définition d'une liste (dite liste de "TYPOGRAPHIE-PROCEDES") permettant une normalisation d'interprétation de certains de ces signes, qui est exploitée pour indexer automatiquement le texte.

Pour le moment, nous ne considèrons que le symbole '"', qui dans le corpus étudié peut être interprété systématiquement comme indicateur de mise en évidence, et introduit donc un concept important. Celui-ci est retenu automatiquement comme indexant l'entité textuelle qui le contient.

3. LE TEXTE EN LANGUE NATURELLE

3.1. Liaison Syntaxe-Sémantique

De nombreux travaux ([COU 77], [COY 72], [DEB 82], [DEW 81], [GUE 82], [FLU 77], [MAN 83], [SYN 70], [VIR 82A]) soulignent le rôle essentiel joué par la syntaxe dans l'analyse du contenu

d'un texte, citons notamment [GUE 82] qui dit:

"La compréhension d'un texte en langue naturelle est composée de deux modes de signifiance: le mode sémiotique et le mode sémantique .

Le mode sémiotique concerne le signe linguistique. Il est en rapport avec les mots de la langue, appelés lexèmes; il doit être reconnu.

Le mode sémantique concerne un contenu et il doit être compris. Il est véhiculé par des parties du discours ou syntagmes; ces syntagmes expriment un contenu et sont donc des CONCEPTS. Dans le texte écrit, ces syntagmes sont représentés par les relations qui existent entre les mots de la langue et ne peuvent donc être déterminés qu'après une analyse syntaxique du texte: la syntaxe établit un lien référentiel entre lexèmes et concepts".

Si la syntaxe concourt à la détermination du sens, il est évident qu'elle ne peut être négligée comme outil d'indexation documentaire des textes.

En analyse documentaire, le but visé est le choix de descripteurs suffisamment précis en sens, qui permettent la sélection ultérieure d'une information pertinente.
Pour arriver à cerner de manière précise et complète le sens d'un descripteur, il faut l'analyser dans son contexte, ce qui nécessite l'extraction d'une partie du discours en langue naturelle.

Dans la méthode proposée, le processus de base pour la reconnaissance des termes d'indexation est l'extraction à partir d'une phrase en langue naturelle des syntagmes les plus longs possibles correspondant à un modèle prédéfini.

Il est clair que plus le syntagme extrait est long, plus l'information qu'il véhicule est complète et précise.

Parmi l'ensemble des catégories de syntagmes, nous nous intéressons plus particulièrement aux syntagmes nominaux, car ils nous paraissent les plus riches et les plus représentatifs quant au contenu du document. Cette limitation est également liée à la conception du langage d'interrogation actuellement envisagé [DEF 84], dans lequel les groupes nominaux jouent le rôle essentiel (les verbes ou auxiliaires ne seront que des mots outils connecteurs).

Dans les structures recherchées, les verbes de la phrase ne sont donc pas retenus même s'ils participent au processus de leur formation au cours de l'analyse.

Le rejet des verbes est justifié par le fait qu'il est rare qu'ils expriment le thème d'une requête documentaire; en général, ils jouent un rôle d'outil pour l'expression de cette requête lorsqu'elle est en langue naturelle. Par ailleurs, les verbes effectivement porteurs d'information peuvent le plus souvent être réduits à une forme substantive sémantiquement équivalente du point de vue documentaire.

Exemple : (câbler, cablâge); (programmer, programmation)

Enfin, on a voulu définir des modèles de syntagmes dont l'analyse soit la moins ambigue possible de façon à garantir un maximum de fiabilité et d'automaticité dans le processus d'indexation. L'exclusion des verbes rejoint également cette préoccupation dans la mesure où ils constituent une source fréquente d'ambiguité syntaxique.

Ces dernières années, plusieurs systèmes existants utilisant des traitements linguistiques évolués [DEB 77], [DEB 82], [FLU 77], [DEW 81], considèrent le syntagme nominal comme élément descripteur du texte. Celà est dû à plusieurs raisons:

- le syntagme nominal est une information suffisante en documentation en ce sens qu'il constitue un compromis entre le fait de garder la phrase complète, ce qui ne serait pas raisonnable, ou de décomposer cette phrase en mots isolés, ce qui reviendrait à une indexation classique.

- L'extraction du syntagme nominal du texte ne nécessite pas des outils syntaxiques complexes; une analyse de surface (cf III), mettant en oeuvre des règles de grammaire simples, permet cette opération en limitant les risques d'ambiguité, souvent rencontrés lors d'une analyse profonde de la phrase (plusieurs interprétations possibles).

- Dans le syntagme nominal, le complément qui suit éventuellement le mot principal permet d'en préciser le sens. On aboutit ainsi à un concept spécifique du concept exprimé par le mot de départ (celui-ci restant par ailleurs disponible en tant que concept générique).

- Le choix de garder le syntagme nominal comme élément d'indexation permet d'éviter l'opération de concaténation (voir indicateurs de rôle et de liaison en III) utilisée en indexation classique par mots-clés, et qui est souvent source d'ambiguité.

- Dans un syntagme nominal, les prépositions établissent des relations qui peuvent exprimer certains rapports (temps, lieu, hiérarchie...) entre les éléments composants (cf IV) tout en affinant l'information véhiculée.

Dans la stratégie proposée, la phase d'analyse syntaxique permet de décomposer la phrase d'entrée en un ensemble de syntagmes nominaux. Chacun d'eux est ensuite transformé en un ensemble de sous-structures appelées groupes conceptuels primaires (notés GCP) correspondant à une forme normalisée des connaissances

reconnues dans le texte. Cette transformation des syntagmes nominaux est nécessaire à la détermination et au stockage optimaux des groupements d'indexation qui seront finalement retenus (cf IV). Enfin, si les termes d'indexation correspondront aux structures des GCP définis ci-après, il importe d'en contrôler le nombre pour éviter une explosion combinatoire qui engendrerait une liste démesurée. La stratégie développée en IV.2 et IV.3 propose d'effectuer cette sélection via une base de connaissances prédéfinie, reflétant l'ensemble des concepts significatifs du corpus, [BCK 83], [BRU 80b], [BRU 81], [BRU 82].

3.2. Définition informelle des GCP

En première analyse, un GCP est considéré comme une sous-structure d'un syntagme nominal; il peut présenter trois formes possibles selon sa taille et par conséquent la quantité d'information qu'il véhicule.

1) Le mot isolé :
Il s'agit d'un substantif (ou nom propre), non qualifié; un syntagme peut être réduit à un substantif lorsqu'il y a risque d'ambiguité dans l'extraction d'une structure donnée (dans ce cas il y a abandon de cette structure au profit du seul substantif) ou tout simplement lorsque le substantif n'appartient pas à un syntagme plus large.

2) Le groupe conceptuel primaire simple :
Il s'agit d'un substantif suivi d'une qualification simple donc de premier niveau.
Généralement, cette qualification est introduite par un adjectif ou par un complément de nom simple.
On distingue deux types de qualifications de ce niveau:

- Qualification indissociable du mot simple:
  il s'agit du syntagme nominal qui en aucun cas ne peut être décomposé sans risque d'introduire une grande ambiguité à l'interrogation, par suite de perte d'information.
  Cette qualification détermine en fait exactement le concept introduit par le mot simple et remédie en partie au phénomène sémantique de polysémie (cf III).
  Elle peut être introduite par un nom propre ou un substantif (EX: 'système expert', 'rapport NORA'...), être également exprimée par une préposition suivie directement d'un substantif sans article (EX: 'imprimante à laser', 'mémoire à bulle', 'manuel d'utilisation').
  Dans certain cas enfin, elle est introduite par un infinitif (EX: 'carte à wrapper', 'machine à écrire').
  Ceci ne peut être généralisé à tous les verbes: en effet, les verbes outils de la langue sont souvent utilisés dans une telle structure sans être, pour autant, porteurs d'une information complémentaire (EX: 'document à voir'; 'contrat à suivre'...).
  En utilisant une classification poussée des verbes [GRO 75], par exemple, on pourrait dissocier les concepts "vrais" des structures à compléments .

- Qualification qui affine le sens du mot isolé mais qui peut être éliminée dans certains cas (voir stratégie de cassure d'un GCP en IV).
  Cette qualification est introduite par un adjectif qualificatif (EX: 'ligne préférentielle) ou par un complément de nom avec article (EX: 'ligne de l'autocommutateur').

3) <u>Groupe conceptuel primaire complexe (GCPC)</u>

C'est toute autre structure plus étendue que les deux précédentes (EX: 'limite des équipements à inclure dans les devis').

Un GCPC peut être extrait tel quel d'un syntagme de la phrase

analysée, ou être le résultat de la connection de deux GCP à l'aide d'une relation binaire tirée du groupe verbal de la phrase (cf IV).
Un GCPC peut véhiculer une quantité d'information importante correspondant à un concept très précis, inversement, sa reconnaissance peut se heurter aux problèmes d'ambiguités; si ce risque est détecté lors de l'analyse, il y a abandon de cette structure complexe au profit des groupes conceptuels primaires simples composants.

On trouvera en VI-1 une définition formelle complète de ces différents types de groupements, et en VI-2 et VI-3 les principes de dérivation des termes d'indexation finaux à partir de ceux-ci.

4. CONCLUSION

La structure logique arborescente du document a permis d'en donner une modélisation en tant qu'ensemble d'entités de structure introduites par des titres. Les informations textuelles proprement dites sont rattachées aux feuilles de cette arborescence. Chacune de celle-ci est transformée, via une analyse morpho-syntaxique (cf IV), en un ensemble de groupes conceptuels primaires qui constituent la représentation des concepts reconnus dans cette entité.
Le passage de ce niveau de représentation aux termes d'indexation finalement retenus (appelés groupes conceptuels finaux, ou GCF), nécessite une étape complémentaire de normalisation de façon à contrôler la croissance de cet ensemble.
Cette seconde phase est fondée sur l'exploitation d'un thésaurus contenant les concepts du domaine couvert. Ceci permet la sélection de GCF qui sont des sous structures dérivées de la structure du GCP initial (cf IV.2 et IV.3).

De ces deux éléments il ressort l'approche unificatrice suivante: tout le document peut être formellement décrit par des outils syntaxiques prenant en compte son aspect structurel global et le sous-ensemble de ses entités textuelles pouvant être représenté dans la structure des GCP.
On peut en déduire une première définition de la relation d'indexation (RI): c'est une relation binaire dont le premier argument est l'ensemble des unités d'indexation (UI) du corpus et le second l'ensemble des GCF extraits de ces unités.
soit:

$$RI \subset UT \times GCF$$

CHAPITRE III

METHODES D'INDEXATION EXISTANTES

1. PREAMBULE

Pour réaliser le processus d'indexation, trois étapes sont indispensables:

- reconnaissance de tous les concepts du document, c'est la phase d'analyse.
- établissement de la relation d'indexation entre cet ensemble de concepts et le document; c'est la phase de sélection et de pondération.
- choix du vocabulaire qui traduit le mieux ces concepts, c'est la phase d'enregistrement.

Les deux premières phases aboutissent, en fait, à la valuation de la représentativité mutuelle entre un document et un concept véhiculé explicitement par un ou plusieurs termes de celui-ci. La troisième phase permet de généraliser cette relation pour l'ensemble des termes qui décrivent le concept dans la langue. Cette phase peut être expliquée ainsi:
En langue naturelle, un concept est décrit par un ensemble de termes du vocabulaire, généralement liés par des relations sémantiques et syntaxiques; il est nécessaire lors de l'analyse ou la recherche des documents, de pouvoir faire référence à une représentation normalisée de ces concepts, quelle que soit la formulation qui en est faite dans les textes ou les requêtes.
Dans la majorité des SRI existants, [AND 75], [GOL 72], [MIS 78], [SAL 71], l'outil linguistique utilisé pour la description de l'ensemble des concepts d'un domaine est le thésaurus.

1.1. Définition et role du thésaurus à l'indexation

Un thésaurus peut être vu comme un ensemble de termes et un ensemble de relations sémantiques et parfois syntaxiques entre ces termes, [SAL 75], [RIJ 79], [BRU 81], [KER 81], [SAL 83].
Ces relations ont généralement une interprétation sémantique universelle: synonymie (pure ou partielle), hiérarchie (généricité, spécificité), causalité, voir_aussi... ou une interprétation contextuelle relative au corpus traité.

Dans un SRI, le rôle essentiel du thésaurus est la représentation normalisée des concepts du domaine, de telle sorte que chacun d'eux puisse être désigné, à l'analyse et à la recherche documentaires, par l'ensemble des termes du vocabulaire qui le décrivent; il constitue le médiateur entre le contenu des requêtes et celui des documents de la base: c'est donc une interface indispensable entre l'indexation et l'interrogation.
Dans les systèmes existants, il est utilisé à des degrés divers au cours de ces deux phases, [AND 75], [FLU 77], [GOL 72], [MIS 78], [SAL 75], [SAL 83].

Pour l'analyse des textes, le thésaurus peut être soit établi préalablement à toute indexation, soit établi au fur et à mesure de l'indexation des documents:

Thésaurus à priori.
Cette approche est celle du langage documentaire contrôlé où le contenu du thésaurus est susceptible de contenir l'ensemble des termes significatifs du domaine couvert: à l'indexation, tout terme inconnu du thésaurus est généralement rejeté.
Cette méthode à l'avantage de préserver une certaine cohérence à l'indexation vu que la représentation des concepts (par conséquent l'établissement des relations sémantiques) est totalement normalisée.

Thésaurus à postériori.

Cette approche est celle du vocabulaire libre où le thésaurus évolue en fonction de l'indexation: les termes qui indexent un document sont incorporés dans le thésaurus et les relations entre ces termes, sont établies à ce moment.

Cette méthode peut être plus souple et moins coûteuse que la précédente.

Elle présente deux inconvénients majeurs: d'une part, le risque de redondance de l'information (surtout si le système ne dispose pas de lemmatiseur pour l'analyse des textes) ; d'autre part, une certaine incohérence dans le cas où les relations entre mots du thésaurus sont établies par différentes personnes.

L'élaboration d'un thésaurus peut être manuelle ou automatique.

La méthode manuelle [GOL 72], [MIS 78], a l'intérêt d'être conçue par des spécialistes qui maitrisent le domaine, et par conséquent peut être efficace quant au contenu du thésaurus (en particulier pour la déduction de concepts non explicites dans le corpus).

Par ailleurs, les relations sémantiques retenues (synonymie, hiérarchie, ...) sont aussi établies manuellement et il est parfois difficile de les définir en concordance avec le contenu effectif du corpus (synonymies partielles, ...), et de donner à cette opération l'exhaustivité nécessaire.

Enfin, les autres inconvénients de cette méthode sont, d'une part son coût (surtout pour la constitution initiale), et d'autre part le risque permanent d'incohérence au niveau du contenu lorsque le corpus évolue.

La construction automatique [BRU 81], [BRU 82], [SAL 75], [SAL 83], de thésaurus consiste en l'établissement d'une relation particulière entre les termes du vocabulaire du corpus traité.

L'établissement de cette relation est effectué à partir de méthodes statistiques basées sur la cooccurrence des couples de termes dans le corpus.

Une relation obtenue de cette manière peut être vue comme une liaison contextuelle entre les termes; elle ne peut évidemment pas avoir une interprétation conventionnelle (synonymie, hiérarchie...), et sa validité est strictement limitée au corpus traité.
L'avantage de la méthode réside dans son faible coût, comparé au thésaurus manuel, et dans le fait que le thésaurus reflète bien le contenu du corpus. L'inconvénient est lié à la relative pauvreté des relations sémantiques ainsi dégagées, et au fait que leur nature précise n'est souvent pas explicitée.

1.2. Problème de précision et de rappel

En introduction, il a été vu que l'efficacité d'un SRI se mesure par les taux de précision et de rappel produits pour les requêtes [SAL 71], [SAL 75], [SAL 83].
Ces deux quantités sont bien sûr fonction de la valuation de la relation d'indexation établie explicitement entre le corpus et les termes d'indexation, mais elles sont très sensibles également à la qualité de la normalisation des concepts dans le thésaurus: souvent les sélections de documents non pertinents ou les pertes de documents pertinents sont dues à une mauvaise interprétation des concepts de la requête, mais aussi à une mauvaise interprétation de ceux-ci à l'indexation.
En fait, les deux premières phases de l'indexation reflètent ce qui existe effectivement dans le corpus, alors que la troisième phase permet de faire des inférences au niveau des termes d'indexation par le biais des relations établies entre les termes du thésaurus.
En particulier, parmi ces relations certaines sont favorables à la précision, d'autres au rappel: la relation contextuelle

établie automatiquement à partir du corpus, peut être favorable à la précision car elle reflète le véritable contenu du corpus (c'est le cas également de la synonymie pure, phénomène rare dans une langue).
Les relations classiques universelles peuvent favoriser le rappel, si elles sont exploitées avec précaution, vu leur caractère général.

2. METHODES STATISTIQUES ET METHODES LINGUISTIQUES

Traité manuellement, le processus d'indexation implique un travail délicat et fastidieux (temps nécessité très important). Du point de vue qualitatif, le grand reproche qui est fait à cette méthode est le manque de normalisation et de cohérence: l'indexation d'un même document peut être différente d'un indexeur à un autre, même si le travail est effectué à partir de listes normalisées de termes d'indexation telles que les thésaurus.

Pour pallier à ces inconvénients, on a tenté de développer des méthodes d'indexation automatique basées sur des critères statistiques et des critères linguistiques, [AND 73A], [AND 73B], [BOO 75], [COO 78], [FLU 77], [MAR 77], [RIJ 79], [SAL 71], [SPA 71], généralement exploités à des degrés divers dans les systèmes.

2.1. Méthodes Statistiques

Les statistiques ont été introduites depuis longtemps en analyse automatique du texte, [LUH 59], car elles fournissent des éléments utiles quant à l'interprétation de certains phénomènes linguistiques.
En documentation automatique, il s'agit, dans les méthodes classiques, d'extraire un élément (généralement un mot) jugé caractéristique d'un texte en fonction de sa fréquence

d'apparition dans ce texte. Des assertions telles que: "la fréquence d'un mot dans un texte fournit une mesure utile de la représentativité du mot dans le texte ", ou "la cooccurrence relative, dans une phrase, de mots auquels ont été affectés des poids de représentativité, est une mesure utile de la représentativité de cette phrase" [LUH 59] sont caractéristiques de ces méthodes très largement employées.

Les procédés statistiques permettent donc d'évaluer, l'importance du mot par rapport au contenu d'un texte en fonction de sa fréquence d'occurrence, en s'appuyant souvent sur des modèles statistiques connus (formule de BAYES, loi binomiale, loi de poisson, valeur discriminante...), [AND 73A], [AND 73B], [BOO 75], [COO 78], [FLU 77], [MAR 77], [SAL 71], [SAL 75], [SAL 83], [SPA 71], [WAL 79], [WUS 83].

Dans les systèmes existants, on distingue deux types d'indexation:

- indexation binaire ou sélective [PAS 72], [MIS 78], [FLU 77]:
  Un terme indexe ou n'indexe pas un texte de document donné. Il est retenu comme terme d'indexation de ce texte si sa fréquence d'ocurrence dépasse un seuil donné (cf III-2-2).

- indexation pondérée [SAL 75], [SAL 83]:
  Pour un terme d'un texte donné, on associe un poids, basé sur la fréquence d'occurrence de ce terme; ce poids reflète l'importance du terme par rapport au texte traité (cf III-2-2).

De plus, le comportement d'un terme est souvent analysé à deux niveaux [FLU 77], [RIJ 79], [SAL 83], [WUS 83]:

- Localement au texte traité:
  La fréquence d'occurrence du terme dans le texte est comparée

aux diverses fréquences d'occurrence des autres termes du texte.

- Globalement dans le corpus traité:
  Le comportement du terme est analysé par rapport à la totalité du corpus, et son importance en tant que terme d'indexation est fonction de diverses mesures évaluant notamment son pouvoir discriminant.

Ces deux niveaux d'évaluation peuvent être interprétés ainsi:

Le niveau local permet de calculer l'importance du terme dans le texte en comparaison avec le reste des termes du texte.

Le niveau global permet de classifier l'ensemble des termes d'indexation, suivant leur pouvoir discriminant par rapport au corpus.

Cette classification permet de définir des catégories de termes; chaque catégorie présente un comportement particulier quant à son importance en tant que terme d'indexation et est traitée d'une manière propre.
Cette classification, basée sur la répartition des termes dans les documents, est réalisée et interprétée différemment dans les différents systèmes:

1) <u>DISSIMILARITE DES DOCUMENTS</u>
   [SAL 71], [SAL 75], [SAL 81], [SAL 83], [WUS 83].
   Les termes ayant un bon pouvoir discriminant, sont les termes qui rendent dissimilaire l'ensemble des documents du corpus ; cette dissimilarité est fonction de la fréquence documentaire (nombre de textes dans lesquels le terme apparaît) du terme (cf III-2-2).

2) DISTRIBUTION DES TERMES DANS LE CORPUS

[COO 78], [RIJ 79].

Dans l'ensemble des termes, on distingue les mots outils de la langue et les mots significatifs qui peuvent indexer les documents.

La distinction entre ces deux sous-ensembles est faite en examinant la répartition des mots à travers l'ensemble des textes des documents :

* mots outils: leur répartition dans le corpus suit une distribution uniforme.
* mots significatifs: leur répartition dans le corpus n'est pas uniforme.

3) AUTRE METHODE

[AND 73A], [AND 73B], [FLU 77].

Un terme ayant un bon pouvoir discriminant est un terme fréquent dans un sous-ensemble restreint de documents, et rare dans le complémentaire de ce sous-ensemble de documents.

2.2. Insuffisances des Méthodes Statistiques

Les méthodes statistiques ne peuvent être utilisées seules car elles ne rendent pas compte des aspects linguistiques d'un texte traité, qui sont très importants pour l'analyse de son contenu: les mots de la langue sont liés par des relations syntaxiques, et sémantiques, et ne pas tenir compte de ce fait conduit nécessairement à des résultats incomplets et imparfaits.

A cet effet R.DEWEZE [DEW 81] remarque: "chacun des mots d'un texte ne porte pas en soi la totalité de sa signification; celle-ci est déterminée à partir des structures syntaxiques et contextuelles de la langue".

En particulier, les problèmes linguistiques suivants situent les limites des méthodes purement statistiques :

1) Normalisation du vocabulaire

Un mot de la langue présente différentes formes grammaticales dans les textes traités. Pour obtenir la fréquence d'occurrence exacte d'un terme, il est donc indispensable de normaliser toutes ses formes vers une forme unique.

Cette normalisation du vocabulaire est effectuée via une analyse morphologique.

2) Extraction de concepts à partir des mots de la langue

En indexation, on s'intéresse à l'extraction de concepts; un concept n'est pas identique à un mot de la langue, il est représenté par un ensemble de mots de la langue mis en relation, ensemble appelé syntagme. L'expression d'un concept n'est généralement pas limitée à un mot de la langue, il est représenté par un ensemble de mots mis en relation à l'intérieur d'un syntagme.

La reconnaissance d'un syntagme ne peut être réalisée qu'à partir d'une analyse syntaxique du texte.

La syntaxe permet également de lever l'ambiguité qui réside au niveau de la morphologie; certains phénomènes d'homographie peuvent être résolus à ce niveau.

3) Relations sémantiques

Les mots de la langue sont donc en relation sémantique (synonymie, antonymie, généricité, causalité...) qui sont souvent source d'ambiguité dans l'analyse automatique de textes. Les relations sémantiques de synonymie et de polysémie sont particulièrement critiques en indexation. Elles faussent le calcul des fréquences si elles ne sont pas pris en compte:

- Synonymie: le calcul de fréquences étant effectué sur le mot et non sur le concept, la fréquence du concept est diminuée (ventilée entre les différents mots synonymes).

- Polysémie: un mot peut représenter des concepts divers, relatifs à des domaines différents; le calcul des fréquences est là aussi faussé si ces divers concepts ne sont pas discernés (concentration des fréquences sur un seul terme qui a en fait plusieurs significations).

4) Mots outils

Un texte est constitué d'un fort pourcentage de mots outils de la langue: ce sont des mots communs à tous les domaines, qui ne sont donc pas significatifs d'un contenu. Ils présentent des fréquences d'occurrence très irrégulières dans les textes traités. S'ils ne sont pas écartés lors de l'analyse, l'interprétation de la mesure d'une fréquence relative perd beaucoup de sa précision. Ces mots outils peuvent être reconnus et éliminés, en partie, par un traitement linguistique: certaines catégories grammaticales correspondent à des classes de mots outils de la langue. Le procédé le plus courant pour les repérer est l'établissement d'un dictionnaire de ces mots, souvent appelé antidictionnaire. Lors de l'analyse d'un texte, la consultation de cet antidictionnaire est nécessaire, et tout mot lui appartenant est éliminé des termes d'indexation.

Enfin, un autre facteur important pour l'efficacité des méthodes statistiques en indexation automatique, est relative à la taille des textes traités: la longueur d'un texte doit être relativement importante pour obtenir des calculs de fréquences (locales, en particulier) significatives.

Dans la plupart des sytèmes existants, l'indexation automatique porte sur le résumé des documents, ce qui n'est pas très

favorable de ce point de vue. Ces dernières années, on s'oriente vers l'analyse automatique des textes complets grâce notamment aux progrès réalisés en matière de capacité de mémorisation et de traitement.
Les méthodes statistiques sont donc favorisées par cette évolution.

2.3. Méthodes Linguistiques

Le traitement linguistique a suscité beaucoup d'intêret, étant donné ses conséquences dans le domaine documentaire [COY 72], [COU 77], [DEB 77], [FLU 77], [DEW 81], [DEB 82], [DES 82], [GUE 82], [MER 82], [VIR 82A], [MAN 83].
Pour tenir compte des aspects linguistiques évoqués plus haut, les concepteurs de systèmes ont donc associé aux méthodes statistiques des outils complémentaires, manuels ou automatiques qui sont de deux types:

- Outils morpho-syntaxiques:
  Ces outils permettent de faire un prétraitement dans l'analyse du contenu d'un texte; ils concernent la normalisation du vocabulaire et souvent, la reconnaissance 'primaire' des concepts du texte.

- Outils sémantiques:
  Ces outils complètent souvent le traitement morpho-syntaxique pour l'analyse du texte; ils sont nécessaires pour la résolution de certaines ambiguités de la langue naturelle.
  Ils permettent d'exprimer les relations contextuelles et sémantiques entre l'ensemble du vocabulaire et de définir ainsi un langage documentaire propre au domaine traité.
  Généralement un langage documentaire est défini par un thésaurus qui exprime chaque concept du domaine par l'ensemble des termes du vocabulaire qui le décrivent (cf III.1).

Parmi les divers procédés linguistiques (mis à part le thésaurus) qui ont été introduits en analyse documentaire, on peut citer les suivants:

(1) Indicateurs de rôle et indicateurs de liaison [MAN 83]:
Par ces procédés, on établit manuellement des liaisons entre les mots-clés d'un même document dans le but d'affiner leur interprétation et de supprimer certaines ambiguités.
Les indicateurs de rôle et de liaison ont pu être utilisés dans plusieurs systèmes grâce à leur simplicité [GOL 72], [SAT 74].
Un indicateur de liaison précise qu'il existe une liaison entre un ensemble de mots-clés d'un même document.

EX: technique-documentation-histoire;

La signification de la liaison n'étant pas précisée, ce procédé ne peut éliminer les ambiguités dues à ses diverses interprétations possibles.

Les indicateurs de rôle permettent d'établir, pour un ensemble de mots-clés d'un même document, une sorte de graphe où sont précisées les connexions qui régissent la structure syntaxique de l'ensemble.
Pour chaque mot-clé de cet ensemble, on précise sa fonction dans cette structure:

EX: histoire-A, enseignement-O, documentation-M;
A: action;
O: objet;
M: moyen;

Mais là aussi, les indicateurs de rôle ne lèvent pas toute l'ambiguité quand la fonction d'un mot-clé peut être interprétée de plusieurs manières par rapport aux fonctions des autres mots-clés.

(2) Analyse morpho-syntaxique:

Les procédés décrits précédemment ne sont pas assez puissants pour résoudre les ambiguités dans une langue naturelle, les concepteurs de systèmes se sont intéressés à des outils plus formels et complets permettant un véritable traitement linguistique des textes des documents [COU 77], [DEB 77], [FLU 77], [PAL 81], [SYN 70].

Ce traitement linguistique comprend deux analyses: l'analyse morphologique qui reconnaît les mots de référence de la langue dans le texte, et l'analyse syntaxique qui détermine les divers syntagmes et leurs articulations.

- Analyse morphologique:

  L'analyse morphologique consiste à segmenter le texte en mots, et à leur associer des renseignements linguistiques qui les identifient comme mots de la langue.

  A chaque chaine de caractères, elle associe des renseignements linguistiques qui l'identifient comme mot de la langue.

  Ces renseignements linguistiques sont en général établis à partir d'un dictionnaire des racines des mots de la langue (qui est conçu partiellement ou complètement a priori) et de modèles de décomposition des mots.

- Analyse syntaxique:

  Elle détermine les diverses structures qui mettent en relation les mots qui composent le texte .

  Ces structures respectent un modèle syntaxique défini a priori pour la langue considérée .

  L'analyse syntaxique est conçue différemment dans les systèmes existants; en particulier elle peut être complète ou partielle.

  Dans le premier cas, elle fournit toutes les structures syntaxiques possibles de la phrase [COU 77].

  Dans le second cas, elle dégage, pour le besoin d'une

application particulière, un sous-ensemble de ces structures syntaxiques.

L'un des premiers systèmes documentaires à avoir mis en oeuvre une analyse syntaxique, est SYNTOL [SYN 70].
Il s'agit d'effectuer une indexation dite relationnelle: dans ce procédé, on définit une syntaxe limitée, fondée sur un ensemble restreint de relations logiques et autant que possible non ambigues. Ces relations sont définies a priori entre les mots-clés.

Dans SYNTOL un mot-clé "primaire" appartient à l'une des catégories suivantes:

-ENTITE (ex: système);
-ACTION (ex: indexation);
-PREDICAT (ex: automatique)

L'analyse d'un texte comporte alors les étapes suivantes:

- une première phase morphologique qui transforme tous les mots du texte en leur racine respective.
- une deuxième phase pour la normalisation du vocabulaire d'indexation: le système dispose d'un dictionnaire établi manuellement des mots-clés "primaires". Au cours de cette phase, il y a remplacement des mots du texte par leur représentant éventuel dans l'ensemble des mots-clés primaires.
- une troisième phase consiste à dégager les structures syntaxiques : pour cela, le système explore un réseau syntaxique entre les mots du dictionnaire d'indexation, établi a priori. Une structure syntaxique entre deux mots du texte, n'est retenue que si elle existe explicitement dans le réseau préétabli.

L'ensemble des structures syntaxiques entre les catégories de mots d'indexation est le suivant:

* relation prédicative, R1:
R1 permet la qualification d'une entité par un prédicat;

EX: système R1 automatique;

* relation associative, R2:
R2 exprime les relations agent/action, action/objet, entité/attribut, tout/partie.

EX: indexation R2 document;

* relation consécutive, R3:
Elle exprime les rapports cause à effet;

EX: somnifère R3 sommeil;

* relation coordinative, R4:
Elle exprime la comparaison ou la coordination entre deux objets;

EX: recherche R4 science;

parmi les quatre relations, seule R4 n'est pas orientée (symétrique).

Ce type de procédé, bien que très riche, est fastidieux quant à sa réalisation et sa maintenance.

3. MISE EN OEUVRE DANS TROIS SYSTEMES OPERATIONNELS TYPES

Nous présentons ici trois systèmes documentaires qui utilisent des méthodes statistiques et des procédés linguistiques pour l'indexation automatique des documents .
Il s'agit du système GOLLEM [GOL 72] dont l'indexation est

assurée par le système PASSAT [PAS 72], du système SPIRIT [AND 73A], [AND 73B], [AND 75], [DEB 77], [DEB 82], [FLU 77], et du système SMART [SAL 71], [SAL 75], [SAL 81], [SAL 83], [WUS 83].
L'objet de cette présentation est, à travers trois logiciels assez représentatifs des méthodes actuellement développées, d'en effectuer une présentation concise mais suffisamment détaillée pour analyser les solutions retenues quant aux principaux aspects méthodologiques évoqués plus haut, et par conséquent de mieux situer notre approche par rapport à cet existant.

-PASSAT:

Ce logiciel a été développé en Allemagne par SIEMENS. Utilisé à l'échelle industrielle depuis de nombreuses années, il a été l'un des tout premiers à intégrer une gestion de thésaurus sophistiquée.

-SPIRIT:

Ce logiciel a été développé en France par C.FLUHR et A.ANDREEWSKY et est en cours d'industrialisation en collaboration avec la société CISI.
Bien qu'encore assez expérimental, c'est certainement le système de recherche d'information le plus avancé actuellement sur le plan des traitements linguistiques.

-SMART:

Ce logiciel a été développé aux USA par G.SALTON et son équipe de Cornell University. Bien qu'étant toujours resté au stade de prototype, les différentes études menées autour de ce projet ont permis de mettre en lumière une grande partie des concepts et des méthodes fondamentales de la recherche d'information, éléments ayant été repris dans beaucoup de travaux ultérieurs.

3.1. Aspects Linguistiques

3.1.1. Analyse Morpho-Syntaxique

1- PASSAT:

Pour l'analyse des textes, PASSAT exploite un dictionnaire de racines des mots de la langue ou mots souches. Ce dictionnaire est établi manuellement.
Chaque racine est suivie d'indicateurs décrivant ses propriétés morphologiques et l'ensemble des relations syntaxiques qu'elle présente avec le reste du vocabulaire. Ceci se traduit par l'association, pour chaque mot souche du dictionnaire, de trois octets d'identification:

- Octet de différenciation du mot souche:
  Il permet de distinguer les homonymes ayant même terminaison.

- Octet de terminaison:
  Il pointe vers la liste des terminaisons possibles du mot souche: genre, nombre, conjugaison de verbes...

- Octet de liaison syntaxique:
  Il pointe vers la liste des compositions possibles du mot souche: mots composés de l'allemand.

2- SPIRIT:

L'analyse morphologique utilise un dictionnaire de racines et un ensemble de règles de morphologie;

L'analyse syntaxique est fondée sur l'utilisation de matrices de précédence binaires, pour la détermination de la structure de la phrase; pour la résolution des ambiguités grammaticales, le système utilise un ensemble de règles syntaxiques acquises par

apprentissage.

L'analyse morpho-syntaxique permet de dégager du texte analysé l'ensemble des syntagmes qui le composent.

- Dictionnaire des mots vides:
  le dictionnaire de mots vides est constitué automatiquement par comparaison d'un ensemble de textes scientifiques avec un ensemble de textes littéraires.

- Reconnaissance des homographes:
  La méthode de reconnaissance des homographes consiste à cocher manuellement dans les index tous les homographes et à réunir ensuite automatiquement les champs sémantiques partiels de ces homographes. Cette opération s'effectue sur un certain nombre de textes prédéterminés .

- Références pronominales:
  Les auteurs de ce système se sont également intéressés au problème de la reconnaissance du référent des pronoms. Cette résolution est très utile à l'application de procédés statistiques aux mots apparaissants dans les textes; la fréquence est largement faussée si on ne prend pas en compte, pour chaque pronom analysé, le concept auquel il réfère .

3- SMART:

Les textes des documents subissent au préalable un traitement morphologique et syntaxique mais l'analyse syntaxique, est réactivée après le critère statistique, pour le regroupement en syntagmes de certaines catégories de termes.

3.1.2. Analyse Sémantique

Pour l'indexation des documents, les trois systèmes considérés utilisent, en plus d'un thésaurus manuel (exploité à des degrés divers (cf III.2.3)), des procédés sémantiques propres.

1- PASSAT:

Pour l'indexation d'un document par un terme PASSAT considère toutes les associations sémantiques avec le reste du vocabulaire que ce terme évoque dans le domaine couvert. Ces associations sont enregistrées dans une matrice, dite de liaisons sémantiques.

La matrice contient l'ensemble des associations sémantiques établies manuellement pour le vocabulaire du domaine traité. Ces associations n'ont pas une interprétation sémantique normalisée et sont obtenues de la manière suivante :
Tous les concepts qui couvrent le domaine sont placés sur deux axes de coordonnées où ils se succèdent dans le même ordre; une racine du dictionnaire est placée à l'intersection d'une abscisse et d'une ordonnée lorsqu'elle présente une parenté sémantique avec le concept en ordonnée et le concept en abscisse.

EX:''libéralisme' sera associé au couple de concepts ('idéologie','libéral');

Une racine peut être associée à différents couples de concepts.

EX: 'fascisme' lié à ('idéologie', 'radical de droite') et à ('dictature', 'Allemagne').

2- SPIRIT:

Les principales relations sémantiques (synonymie, généricité) sont reconnues automatiquement dans SPIRIT:

D'une manière générale, la reconnaissance d'une relation sémantique entre un concept ci et un concept cj, est fondée sur :

* la détermination du champ sémantique si du concept ci.
* la détermination du champ sémantique sj du concept cj .
* une relation ensembliste sur si et sj .

On a alors :

a- Synonymie: $si \cap sj \simeq si \simeq sj$: le champ sémantique de Ci est voisin du champ sémantique de Cj ; Ci et Cj peuvent être confondus comme synonymes .

b- Généricité:
$si \cap sj = sj \dashrightarrow sj$ si : Cj est terme spécifique de Ci .

c- Parenté sémantique:
$si \cap sj \neq 0 \dashrightarrow$ Ci et Cj sont dans une relation de parenté.

3- SMART:

Dans SMART une relation de voisinage sémantique est établie automatiquement entre l'ensemble de T termes du domaine couvert.

Pour chaque terme i, on considère son degré d'association avec chacun des autres termes du vocabulaire. Cette association est basée sur la fréquence de cooccurrence des termes dans les documents.

Une matrice de dimension (T x T), dite matrice associative permet d'enregistrer pour chaque couple (i, j) du vocabulaire, leur degré de liaison mutuelle.

Cette liaison est évaluée par un coefficient de similarité entre les termes i et j; elle se calcule ainsi:

$$S(i, j) = \frac{\sum Fik.Fjk}{\sum_k Fik^2 + Fjk^2 - \sum_k Fik.Fjk}$$

où:

N: nombre total de documents considérés.

Fik: nombre d'occurrence du terme i dans le document k.

Le coefficient de similarité prend sa valeur dans l'intervalle [0,1]:
il est égal à 0 si les termes i et j ne cooccurrent dans aucun document; dans le cas contraire, sa valeur se rapproche de 1.

3.2. Aspects Statistiques

1- PASSAT:

Le critère statistique est basé sur la loi de ZIPF: les fréquences d'occurrence des mots d'un texte analysé, sont rangées par ordre décroissant et sont comparées aux fréquences d'occurrence des mots d'un texte de référence ou composite.

Tout mot du texte n'appartenant pas au dictionnaire des mots souches ou à la matrice d'associations est rejeté.
Après ce rejet, les éléments susceptibles d'indexer le texte, sont issus des trois catégories suivantes:

- les mots du texte possédant des associations dans la matrice des liaisons sémantiques.
- les mots formant un mot composé de la première catégorie.
- les divers concepts de la matrice qui sont associés aux mots de la première catégorie.

Un traitement sémantique (cf III.2.3) permet de désigner les descripteurs effectifs du document.

2- SPIRIT:

Pour les auteurs de ce système, l'indexation d'un document est la détermination d'une fiche d'identité différentielle de ce document (identité signifie la représentation du contenu sémantique partiel propre au document).
Cette identité différentielle peut être définie en termes ensemblistes et statistiques :

- Définition ensembliste : l'index d'un document, est constitué par le "complémentaire, dans chaque document, de la réunion des intersections prises deux à deux" autrement dit on ne peut avoir un index identique ou presque pour deux documents du corpus .

- Définition statistique :
  L'intensité P avec laquelle un document d est caractérisé par un concept c est d'autant plus grande que c est fréquent dans d et rare dans d<>di. Ceci peut être modélisé par la formule de BAYES qui s'écrit (où n est le nombre de document du corpus):

  $$P(di \mid ck) = P(ck \text{ dans } di)/\sum_j P(ck \text{ dans } dj) \;;$$

  Pour décider si ck peut indexer di, il suffit de comparer la valeur P(di si ck) à la valeur moyenne Pk, calculée sur tous les di:

  $$Pk = (1/n) \sum_j P(dj \mid ck) \;;$$

3- SMART:

Dans SMART chaque document du corpus D est représenté par un vecteur document vec= t1,...,tT de dimension T.
L'élément tij représente le poids du terme tj par rapport au

document di (fréquence d'occurrence de tj dans di).
A partir de ces éléments, on définit un espace documentaire composé des divers vecteurs documents .
Le but de l'indexation est une redéfinition de ces vecteurs documents telle que l'espace documentaire soit le plus 'dispersé' possible; l'hypothèse étant qu'un espace documentaire très dense ne peut être utile à la recherche documentaire puisque la distinction entre les divers di dans D est faible.
La proximité entre deux vecteurs documents relatifs respectivement à di et dj est définie par un coefficient de similitude. Le coefficient de similitude entre di et dj est petit lorsque les vecteur veci et vecj sont éloignés l'un de l'autre dans l'espace documentaire .

L'importance d'un terme ti par rapport au corpus est, par hypothèse, proportionnelle à l'écart qu'il produit dans l'espace documentaire entre chaque couple (di,dj) dans D. Cette importance se mesure par 'la valeur discriminante', DVi, de ti qui donne la différence de compacité de l'espace documentaire entre l'existence et la non existence du terme ti dans l'ensemble T.

L'écart entre deux vecteurs documents est mesuré par le cosinus de l'angle qu'ils forment dans l'espace documentaire.

Le calcul de la valeur discriminante pour chaque élément de T permet de subdiviser ce dernier en trois catégories :

- DVi>0: ensemble des bons descripteurs qui donnent un espace documentaire dispersé.
- DVi=0: ensemble des descripteurs pauvres qui laissent inchangé l'espace documentaire.
- DVi<0: ensemble des mauvais descripteurs qui donnent un espace documentaire très dense.

Les termes ti tels que DVi<=0 ne sont pas rejetés: ils subissent un traitement spécifique (cf III.2.3) de telle sorte que la valeur discriminante qui leur est associée devienne positive.

G.SALTON a établi une correspondance entre la valeur discriminante d'un terme et sa fréquence documentaire, FDOC (S1, S2, S3 étant des seuils empiriques):

- S1<FDOCi<S2: DVi est positive et ti peut être un bon descripteur .
- FDOCi<S1: DVi est nulle et ti est un pauvre descripteur .
- FDOCi>S2: DVi est négative et ti est un mauvais descripteur .

3.3. Aspects Sémantiques

1- PASSAT:

1) Exploitation de la matrice d'associations sémantiques:
A chaque élément des trois catégories de termes susceptibles d'être descripteurs, on affecte un poids de la manière suivante:
   - à un concept de la troisième catégorie, est affectée sa fréquence d'associations sémantiques avec les mots du texte.
   - à un mot de la première ou seconde catégorie, est associée la somme des poids des concepts de la troisième catégorie auxquels il est lié.

   Les concepts retenus comme indexant le texte, sont ceux dont le poids est supérieur à un seuil prédéterminé, relatif au texte composite; il s'agit donc d'une indexation avec sélection.

2) Exploitation du thésaurus:
Tout document indexé par un terme est automatiquement indexé

par l'ensemble de ses synonymes, extraits du thésaurus. En fait parmi cet ensemble de termes équivalents, un seul est désigné comme descripteur principal, les autres étant des descripteurs secondaires: toute référence aux termes secondaires renvoie au terme principal dans le thésaurus.

2- SPIRIT:

Pour pallier à la fausse interprétation des fréquences, le système utilise une fonction de poids sémantique:

Le calcul de la quantité $P(di \mid ck)$ est en accord avec la théorie de l'information qui dit qu'un concept apporte d'autant plus d'informations qu'il se rencontre rarement dans le corpus. Pour remédier à une fausse interprétation de cette quantité quant aux termes rares mais insignifiants pour le domaine, le système utilise des fonctions de poids sémantique qui évaluent "l'importance du concept par rapport au corpus".
La fonction de poids sémantique d'un concept peut être calculée soit par l'utilisation des relations sémantiques de ce concept dans le thésaurus, soit par l'utilisation de la formule de l'entropie ou de SHANNON.

- Utilisation de la formule de SHANNON:

  $$H(ci)=-\sum_j P(dj|ci) \times \log_2 P(dj|ci);$$

  Si ci est très rare H(ci) tend vers 0 et si ci est uniformément réparti sur l'ensemble des documents, $H(ci)=\log_2 n$.

- Utilisation d'un dictionnaire de spécialité ou d'un thésaurus:

  Si on dispose d'un dictionnaire de spécialité, on peut supposer que la spécificité d'un terme est d'autant plus faible qu'il participe directement ou indirectement à la définition d'autres termes. La fonction de poids sémantique

peut être, dans ce cas calculée de la manière suivante (où ni est le nombre de mots définis par le concept i et m le sup de ni):

$$fi = m - ni + 1;$$

3- SMART:

Les termes pour lesquels leur valeur discriminante est nulle sont des termes trop spécifiques du domaine : ils ne peuvent être rejetés car leur suppression affecte la précision en rendant impossible la discriminance entre document.

Ces termes rares sont élargis par des classes de descripteurs voisins, extraites d'un thésaurus établi manuellement.

Les termes ti pour lesquels DVi est négative sont les termes génériques du domaine; leur suppression détériore le taux de rappel. Dans le système, ces termes génériques sont regroupés en syntagmes de descripteurs pour lesquels la fréquence documentaire est 'raisonnable'; la valeur discriminante qui leur est associée est donc plus importante .

3.4. Conclusion

Passat est l'un des premiers systèmes opérationnels à avoir utilisé des méthodes statistiques pour l'indexation automatique des documents; cependant les critères statistiques utilisés sont assez figés: un terme est examiné localement dans un texte et l'analyse de ce texte n'est pas confrontée à l'ensemble du corpus.
La référence vers un texte composite parait également contraignante: son élaboration manuelle pour un domaine précis nécessite un travail délicat et reste assez subjective.

Un aspect original dans PASSAT est l'établissement des associations sémantiques entre les termes du vocabulaire.
Plutôt que de définir des relations sémantiques conventionnelles et générales (synonymie, généricité...), qui n'expriment pas toujours ce qui existe effectivement dans un corpus, la matrice d'associations décrit des relations contextuelles, qui permettent plus précisément de cerner les thèmes traités dans le corpus.
L'inconvénient majeur à ce niveau, est l'établissement manuel de la matrice d'associations sémantiques.

Dans notre système, les relations contextuelles entre termes sont établies automatiquement, et décrivent le contenu du corpus traité (cf IV).

Dans SPIRIT, les critères statistiques mis en oeuvre permettent une certaine classification des termes d'indexation et leur évaluation s'effectue sur tout le corpus; mais dans ce système, il s'agit d'une indexation binaire et ceci semble assez contraignant pour une mise à jour de l'indexation quand le corpus évolue.
La construction automatique des relations sémantiques (synonymie, généricité...) paraitrait séduisante mais en pratique, les résultats obtenus n'ont pas été très satisfaisants pour l'instant.

L'aspect le plus intéressant dans SPIRIT est le traitement syntaxique que subissent les textes analysés: l'analyse semble assez complète et le point original, à ce niveau, réside dans l'apprentissage automatique des règles syntaxiques.

Dans notre système, l'analyse morpho-syntaxique prévue est largement inspirée des méthodes utilisées dans SPIRIT (cf IV).

Dans SMART, la notion de dispersion de l'espace documentaire et sa redéfinition permet une véritable classification des termes d'indexation et semble intéressante pour les besoins précis de la recherche documentaire.
Dans ce système, il s'agit d'une relation d'indexation pondérée dont la définition tient compte des deux aspects fondamentaux d'une bonne recherche documentaire: le rappel et la précision. Les termes génériques du domaine subissent un traitement syntaxique de telle sorte que leur utilisation favorise le rappel; les termes spécifiques subissent un traitement sémantique et leur utilisation favorise la précision.

Dans notre proposition nous avons retenu la solution d'une indexation pondérée, permettant également une classification des termes de même type que celle de SMART (cf V).

Chacun de ces logiciels présente donc à la fois des aspects méthodologiques intéressants mais aussi des imperfections souvent liées à une grande complexité de mise en oeuvre de certaines fonctions. Notre approche a consisté à essayer de reprendre certains points que nous estimons très positifs dans ces différents systèmes, et à les intégrer dans une stratégie globale d'indexation des documents. Au delà de cet aspect de notre étude, nous avons tenté d'introduire notre propre apport à la solution du problème, en essayant notamment de pousser plus loin le degré d'automatisation des tâches liées à l'indexation des documents.

CHAPITRE IV

STRATEGIE D'UTILISATION DES OUTILS LINGUISTIQUES

Comme indiqué dans le chapitre II, la construction des termes d'indexation est fondée sur deux phases de traitements linguistiques:

- L'analyse morpho-syntaxique qui extrait l'ensemble des Groupements Conceptuels Primaires (GCP) d'une phrase d'entrée, et qui constitue une première ébauche dans la détermination des concepts du domaine couvert (cf II).

- L'exploitation du thésaurus qui permet de sélectionner les groupements finaux d'indexation (GCF); ceci se faisant par le biais d'une comparaison entre les GCP et les concepts retenus dans le thésaurus.
  Nous développons dans ce chapitre ces deux phases essentielles de notre stratégie d'indexation automatique qui conduisent à la construction de la relation d'indexation. L'aspect valuation de cette relation sera détaillé dans le chapitre V.

1. ANALYSE MORPHO-SYNTAXIQUE

Il est important de distinguer dans ce qui suit entre l'analyse syntaxique d'une phrase (la reconnaissance des groupes nominaux), et la production des structures qui seront retenues comme termes d'indexation.
Il n'y a correspondance entre ces deux notions que dans certains cas simples.
L'analyse correspond à une syntaxe prédéfinie des groupes nominaux et de certains connecteurs existant entre eux (verbes, auxiliaires, prépositions). Cette syntaxe est forcément

limitative pour des questions de performances, et pour éviter au maximum les traditionnels problèmes d'ambiguité qui pourraient interrompre le processus automatique. Les groupes produits constituent le standard adopté pour la représentation des termes d'indexation. La production des GCP pourrait donc être modélisée par une grammaire transformationnelle classique.

Il est clair que plus le GCP produit est long, plus l'information véhiculée est complète et précise; inversement, il risque de n'avoir qu'une faible représentativité quant au contenu global du texte. Une stratégie de limitation est donc nécessaire pour assurer l'équilibre classique entre précision et rappel. Nous proposons de fonder cette stratégie sur l'utilisation d'un thésaurus normalisateur et sur un processus de cassure des GCP jugés trop longs (cf IV.3).

1.1. Formes d'un GCP

Un GCP constitue donc la représentation choisie pour un groupe nominal; il peut être construit de deux manières possibles:

- Il peut être identique à un groupe nominal trouvé dans la phrase d'entrée (Exemple: 'ligne d'abonnés').

- Il peut être construit par la connection de deux syntagmes nominaux de la phrase. Généralement cette connection est réalisée par une préposition qui précède ou qui suit le groupe verbal de la phrase.

On distingue trois formes possibles de GCP. Nous reviendrons plus loin sur ces cas particuliers, correspondant à des niveaux de complexité croissants. Cette distinction est utile lors de la reconnaissance de ces éléments dans le thésaurus (cf IV.2) et lors de l'application des critères statistiques (cf V). Les structures syntaxiques respectives de ces catégories sont les

suivantes:

1) Le mot isolé

C'est un GCP réduit à un substantif.

MI::= SUBSTANTIF.

2) Le groupe conceptuel primaire simple (GCPS)

Il s'agit d'un substantif dont la qualification simple est juxtaposée dans le texte.

La syntaxe d'un GCPS est la suivante:

GCPS::= SUBSTANTIF G2

G2::= SUBSTANTIF / ADJECTIF / PREP G3

G3::= SUBSTANTIF / VINF

PREP::= 'préposition'

VINF::= 'verbe infinitif'

EX:'ligne préférentielle'; 'ligne d'abonnés'; 'équipement à inclure'.

Remarque:

Comme il a été souligné en II-2, on distingue deux catégories de GCPS pour la structure 'SPS' suivant la forme du groupe nominal d'origine: présence ou absence d'un article précédent le second substantif.

3) Le groupe conceptuel primaire complexe (GCPC)

Un GCPC peut être construit de deux manières différentes:

a) Il peut être extrait tel quel du texte analysé.
(EX: 'limite des équipements à inclure dans devis').
Dans ce cas il correspond à la syntaxe suivante:

GCPC1::= S1 [ADJECTIF]* [PREP S2]*

S1::= SUBSTANTIF/ SUBSTANTIF SUBSTANTIF

S2::= S1 [ADJECTIF]* VINF

On vérifie aisément que cette syntaxe recouvre le cas des mots isolés (MI) et celui des GCPS précédemment définis.

b) Il peut être obtenu par la connection de deux GCP plus restreints:
A partir des GCP du type précédent, on construit des structures plus complexes, en regroupant un couple de GCP connectés par une relation particulière.
Pour le moment, on considère seulement trois types de relations extraites des groupes verbaux de la phrase:

- La relation prépositionnelle (RELP):
  Elle est introduite par une préposition qui suit ou précède directement un groupe verbal dans la phrase.

  Ce type de relation peut exprimer un rapport de temps, de lieu, de moyen... entre deux GCP.

  EX: l'oiseau chante sur l'arbre => oiseau SUR arbre.

- La relation générique (AUX):
  Elle est introduite par l'auxiliaire "être". Elle peut exprimer une relation hiérarchique entre deux GCP.

  EX: l'oiseau est un animal => oiseau AUX animal.

  Pour le moment, on ne considère que cet auxiliaire car en ce qui concerne l'auxiliaire avoir, cette propriété n'est pas toujours vraie étant donné que cet auxiliaire exprime plus souvent une propriété de possession (EX: l'homme a un chapeau).

- La relation de proximité (PROX):
  Cette relation introduit une connection entre un GCP et son complément d'objet direct.

  EX: l'enfant mange la pomme => enfant PROX pomme.

Ces trois relations doivent refléter le plus possible la structure de la phrase d'entrée; lors de leur établissement, si il y a ambiguité non résolue au niveau de la phrase ou indécision dans le choix de la relation (lorsqu'un groupe verbal est délimité par deux prépositions, par exemple), il y a abandon de la structure correspondante, et retour au niveau des GCPCl. Une étude ultérieure des propriétés de ces relations (transitivité, symétrie...) pourrait permettre d'effectuer des inférences lors d'une recherche dans la base documentaire, dans le but d'augmenter le rappel [DEF 84].

Un GCP correspond en définitive à la syntaxe suivante:

GCP::= GCPCl [R GCPCl]
R::= RELP/AUX/PROX
RELP::= 'préposition'
AUX::= 'être'
PROX::= déduite.

1.2. Outils linguistiques envisagés

Les trois catégories de GCP sont extraites lors d'une analyse partielle du texte. A cet effet, on envisage d'utiliser un analyseur morpho syntaxique de la langue française, dont l'étude est menée en parallèle [PAL 81].
Cet analyseur a deux composantes:

1) L'analyse morphologique

Elle exploite un dictionnaire de morphes, avec leurs caractéristiques propres, et un ensemble de règles morphologiques.
La structure de ce dictionnaire est inspirée de la méthode de M.KAY [PAL 81] qui présente deux avantages.

* un gain en place considérable: le dictionnaire est organisé en arbre et l'information y est complétement factorisée (EX: sous forme de liste, cette factorisation donne pour les deux morphes, 'bois' et 'boite', (bo(i(s,te))).

* Un gain en temps d'exploration: l'analyse d'une chaine s'effectue caractère par caractère et sans retour arrière.

Une caractéristique très importante de l'analyse morpho-syntaxique projetée, est l'enrichissement automatique du vocabulaire (donc du contenu du dictionnaire), par déduction automatique des racines et des propriétés morphologiques des mots nouveaux. Cet enrichissement a été restreint aux mots réguliers parmi les classes suivantes: verbes, substantifs, adjectifs, adverbes.
Cet enrichissement ne peut être mis en oeuvre que si les mots irréguliers et les mots outils ont été préalablement chargés dans une phase d'initialisation du dictionnaire. On a donc chargé les deux classes de mots suivantes:

- Les classes fermées de catégories du français (articles, pronoms...); ces classes fermées introduisent des mots vides pour le domaine mais sont nécessaires pour l'analyse syntaxique.
- Les mots irréguliers, dont une liste assez complète a pu être établie à partir de diverses sources.

Si on admet que les mots irréguliers constituent principalement un héritage linguistique, on peut considérer que cette classe est également fermée, une fois leur liste donnée. En effet, les mots nouveaux de la langue ont toujours un comportement régulier.
L'enrichissement automatique du dictionnaire a l'intérêt de permettre la poursuite de l'analyse lors de la rencontre d'un mot inconnu du dictionnaire ce qui est fondamental étant donné le grand nombre de textes traité. D'autre part, il assure une certaine cohérence dans le système puisqu'il n'y a pas d'intervention manuelle pour lever des ambiguités.
Signalons, pour terminer, que la déduction de la classe et de la racine d'un mot nouveau s'opère grâce à un parallélisme entre analyse morphologique et analyse syntaxique.

2) <u>L'analyse syntaxique</u>

La méthode retenue est inspirée des travaux de ANDREWSKY et C.FLUHR (cf III-2).
Cette analyse se décompose en deux phases: résolution des ambiguités grammaticales du français et extraction de la structure de la phrase analysée.
Pour résoudre les ambiguités grammaticales, l'analyse utilise un certain nombre de règles binaires et ternaires qui décrivent les relations positionnelles entre catégories grammaticales.
Il y a apprentissage de ces régles à partir d'un texte initial résolu manuellement.
Notre objectif n'étant pas de procéder à des analyses complètes de phrases, mais à reconnaître les groupes dont les GCP puis les GCF sont dérivés, une adaptation de cette méthode est en cours d'étude pour la limiter à nos besoins et en augmenter les performances.

2. ROLE DU THESAURUS

L'exploitation du thésaurus permet la définition des éléments finaux d'indexation; ceci s'effectue par le biais d'un pattern-matching entre les GCP, résultats de l'analyse syntaxique, et les concepts enregistrés dans le thésaurus.
A l'issue de cette opération, chaque entité de structure est reliée à un ensemble de termes d'indexation stockés dans le thésaurus, et appelés groupes conceptuels finaux (GCF).

Le thésaurus [BRU 80A], [BRU 80B], [BRU 81], [BRU 82], est actuellement constitué d'un ensemble de groupes de termes liés par des relations sémantiques contextuelles, représentant chacun globalement un concept important. La seule propriété qui nous intéresse directement dans le processus du pattern-matching avec les GCP, est que ces groupes correspondent à un sur ensemble de la syntaxe des GCP. En d'autres termes, il n'existe pas de groupe dans le thésaurus qui ne puisse être lui même décomposé selon la syntaxe propre aux GCP. Le pattern-matching entre un GCP et le thésaurus se résume donc essentiellement en un test d'inclusion du GCP dans un groupe.

2.1. Structure actuelle du thésaurus dans l'application

Ce thésaurus est construit automatiquement à partir d'un traitement séparé du corpus [BCK 83], [KBC 83]. L'objectif visé lors de cette opération sémantique contextuelle, est l'établissement automatique d'une relation entre les termes du vocabulaire du corpus traité, à partir de méthodes statistiques basées sur la cooccurrence des termes dans les textes.
L'hypothèse sous-jacente est que deux termes qui apparaissent fréquemment dans une même phrase ont de fortes chances d'être en relation sémantique et également de concourir à la définition d'un concept du domaine. Il est à souligner que les informations

ainsi obtenues sont strictement limitées au corpus traité, mais en reflètent aussi très fidèlement le contenu.
La construction d'une relation entre deux termes du corpus tient compte de la proximité géographique de ces termes dans la phrase, de la fréquence d'occurrence de ce couple de termes ainsi que de la catégorie grammaticale respective des deux termes.

Cette relation binaire est actuellement construite en deux phases:

La première phase est l'extraction, pour tout mot lemmatisé du texte, de l'ensemble de ses coordonnées définies comme un numéro de phrase et un numéro de mot dans la phrase. Ces informations permettent la prise en compte de l'éloignement des mots dans le calcul de la mesure d'association des couples de mots.

La deuxième phase consiste en l'établissement d'une matrice terme-terme dans laquelle à chaque couple de termes (ti, tj) du vocabulaire lemmatisé, on associe une liaison prenant sa valeur dans [0, 1] et calculée à partir des coordonnées respectives des termes ti et tj ainsi que de la fréquence de cooccurrence de ce couple dans une même phrase.

A ce niveau, il a été également nécessaire de tenir compte de la catégorie grammaticale des mots pour le calcul de cette distance.
Un graphe de définition des seuils de distances entre catégories a été établi a priori dans le but de:

- Restreindre l'évaluation des liaisons à certains couples de catégories grammaticales jugées significatives (intervenant dans la définition des GCP); ceci permet également de réduire la taille de la matrice et d'en accélérer considérablement la construction.

- Considérer des seuils de distance admissibles, propres à chacun de ces couples de catégories; ceci permet d'interpréter convenablement les résultats.
  Par exemple la distance entre un substantif et un adjectif est limitée à neuf, seuil au delà duquel on considére que ces deux mots ne sont pas syntaxiquement liés (au sens des GCP).

Les différents types de liens considérés sont les suivants:

- catégories grammaticales: substantif et nom propre (S), adjectif (A), verbe (V) et préposition (P).

- seuils de distance entre ces catégories:
  S-S = 20;
  S-A = 9;
  S-V = 20;
  S-P = 5;
  A-A = 8;
  A-P = 5;
  A-V = 9;
  V-V = 20;
  V-P = 5;

La mesure est symétrique et prend sa valeur dans [0, 1]; par conséquent la matrice terme-terme est symétrique. En raison des seuils de distance imposés elle est également très creuse.

A partir de cette relation binaire, la deuxième étape pour la construction du thésaurus consiste à regrouper en classes les sous-ensembles de termes fortement liés.
L'hypothèse faite ici est que ces groupes recouvrent des concepts importants du domaine couvert par le corpus.

Ces groupes sont des sous-graphes complets maximaux (aussi appelés cliques) extraits du graphe global de la matrice

d'associations. De ce fait chaque terme de la clique présente une liaison strictement positive avec chacun des autres termes de cette clique. A chacune d'elles est associé un poids [0, 1], fonction de ces diverses liaisons et qui représente la force de cohésion de ce groupe.

Remarques.

- L'évaluation de la mesure selon les critères statistiques de fréquence de cooccurrence et de distance, n'est appliquée qu'au texte. En ce qui concerne les entités textuelles de type 'titres pleins' notamment, la procédure est différente car on leur accorde une importance particulière (cf I). Le processus d'évaluation consiste alors simplement à affecter la valeur maximale 1 de la mesure à tout couple de termes trouvé dans ces entités, sans limitation de distance. On est alors certain de retrouver intégralement ces groupements importants dans des cliques.

- Une clique n'étant par construction, pas ordonnée (il s'agit d'une parenté sémantique entre mots), elle peut correspondre à un ensemble de GCF par instanciation d'un modèle syntaxique et des mots qu'elle contient (EX: (histoire, de, enseignement) peut donner "enseignement de l'histoire" et "histoire de l'enseignement"). Pour différencier les divers GCF contenus dans une clique, la détermination d'une fonction est nécessaire. Cette fonction est relative à l'ordre des mots choisis arbitrairement pour la clique et permet d'extraire ou de retrouver un GCF donné dans la clique.

- L'ensemble des cliques est enregistré par ordre croissant de taille (cardinal du groupe), et selon l'ordre croissant de leur trace pseudo-syntaxique (cf ci-dessous). On remarque qu'il peut exister des cliques unitaires (un mot isolé), elles correspondent à des mots pour lesquels aucune liaison

contextuelle intéressante (au sens des GCP) n'a été trouvée dans la matrice des liaisons.

Dans l'état actuel de son élaboration, le thésaurus est donc un ensemble de cliques représentant un groupe de termes fortement liés. Par construction, chacun de ces groupes peut être décomposé selon la syntaxe des GCP. Si on considère que chacun de ces groupes représente un ou plusieurs concepts significatifs tirés du corpus, l'ensemble ainsi obtenu est plus proche de la notion classique de base de connaissances que de thésaurus, tels qu'ils ont été présentés jusqu'à présent. Des développements sont en cours pour enrichir cette base de données sémantiques déduites de cette première analyse (mise en évidence de synonymies, de relations génériques ou spécifiques), qui permettront d'affiner encore plus le processus d'indexation.

2.2. Trace pseudo-syntaxique d'une clique

A partir des catégories grammaticales de l'ensemble des termes d'une clique, on peut déduire une trace dite pseudo-syntaxique, qui est la concaténation de ces différentes catégories grammaticales des mots composants.
Dans un but de normalisation, ces traces sont ordonnées selon les catégories grammaticales rencontrées.

Parmi ces traces, certaines peuvent correspondre à des modèles syntaxiques valides au sens de la structure du GCP alors que d'autres en définissent des sur ensembles.
On peut ainsi distinguer deux catégories de cliques, en fonction de leur trace:

- Une trace peut correspondre, à l'ordre près, à un ou plusieurs modèles syntaxiques valides.
  Exemple:
  la trace 'SSP', qui correspond au modèle 'SPS' de groupe.

la trace 'SSAP', qui correspond aux modèles 'SAPS' et 'SPSA' de groupes.

D'une telle trace, on peut bien sûr également extraire des sous-modèles syntaxiques valides; dans 'SPS' par exemple, le seul sous-modèle valide est 'S', alors que pour 'SAPS' on aurait les sous-modèles 'SA', 'SPS', 'SS', 'S'.

- Une trace qui correspond à un sur ensemble de modèles de GCP: on peut en extraire un ou plusieurs modèles valides.
  Exemple:
  La trace 'SSAPP' qui englobe les modèles 'SPSA', 'SAPS, ainsi que les sous-modèles qu'on peut en dériver.

Il ressort de cette analyse de la trace pseudo-syntaxique des cliques que ces groupements correspondent bien à un sur ensemble de la syntaxe des GCP.

2.3. Pattern-matching entre un GCP et le thésaurus

L'opération qui nous intéresse ici est la validation des GCP construits à partir du texte, en tant que GCF.
Le pattern-matching consiste en fait à sélectionner la première clique qui contient le GCP considéré; on a vu dans le chapitre ci-dessus que cela revient à vérifier l'inclusion du GCP dans une clique donnée. Cette inclusion peut être immédiate dans le cas où tous les couples mot-catégorie du GCP se retrouvent dans une clique; il peut également se produire le cas où cette inclusion n'est pas réalisée, ou partiellement réalisée. Cela correspond au fait que le GCP considéré ne correspond pas à un concept important du domaine. Il faut alors envisager une stratégie de fragmentation de façon à réutiliser ce groupe à un niveau sémantique moins précis.
Un GCP donné peut être inclus dans plusieurs cliques du thésaurus; il est cependant impératif de sélectionner, pour

chacun d'eux, une clique qui en devienne le représentant unique lors de traitements ultérieurs. Lors de sa première apparition, le terme d'indexation est enregistré via cette clique, et toutes les informations liées à la relation d'indexation propre à ce terme seront centralisées autour de cette représentation unique. Compte tenu de ces remarques préliminaires, et des propriétés respectives des GCP et des éléments du thésaurus examinés dans la section précédente, on peut définir comme suit le processus de pattern-matching retenu:

- Un GCP qui correspond à l'ordre près à un élément du thésaurus, devient un groupe conceptuel final (GCF) et est considéré comme élément d'indexation.

- Dans le cas contraire, il est fragmenté (selon des critères syntaxiques) en sous groupes qui sont à leur tour comparés au contenu du thésaurus. Ce processus de décomposition est poursuivi tant que les différents éléments obtenus ne correspondent pas au critère de correspondance entre un GCP et un élément du thésaurus (inclusion). La décomposition peut être poussée jusqu'au mot isolé. A la fin du pattern-matching on a donc retenu comme termes d'indexation les syntagmes de plus haut niveau issus du GCP initial et compatibles avec la norme imposée par le thésaurus.
  Le vocabulaire du thésaurus étant un sous-ensemble du vocabulaire du corpus, il peut arriver qu'un ou plusieurs termes du GCP soient inconnus du thésaurus.
  Dans ce cas, deux traitements sont possibles:

- Tous les termes du GCP sont inconnus du vocabulaire des cliques: le GCP est rejeté.

- Un sous-ensemble seulement des termes du GCP est inconnu du thésaurus: dans le cas où ce sous-ensemble contient tous les substantifs qui forment le GCP, ce GCP est rejeté. En effet,

comme toute sous-structure de GCP (d'après la syntaxe d'un GCP) contient au moins un substantif, quelque soit la décomposition de ce GCP, le pattern-matching échouera.
Si ce n'est pas le cas, le GCP est décomposé en sous-structures de plus bas niveaux (GCPS, mot isolé). On relance le pattern-matching sur les sous-structures correspondant à des termes connus du thésaurus.

3. CASSURE SYNTAXIQUE D'UN GCP

La décomposition d'un GCP s'effectue selon des critères syntaxiques de façon à fractionner le concept initial en entités cohérentes, pouvant à leur tour exister dans le thésaurus.
La décomposition syntaxique d'un groupe conceptuel primaire est fonction de sa catégorie (mot isolé, GCP simple ou GCP complexe) et peut être récursive. Elle échoue lorsqu'elle n'aboutit à aucune sous-structure (ceci est évident pour le mot isolé, par exemple); dans ce cas le GCP initial est rejeté en tant que terme d'indexation possible.

Le processus de décomposition s'appuie sur un réseau syntaxique établi a priori (appelé réseau statique), et un réseau dynamique, engendré durant le traitement du GCP.

Le réseau statique permet d'orienter la stratégie de décomposition vers les sous-structures syntaxiquement correctes du GCP.

Le réseau dynamique permet notamment de résoudre les liaisons substantif-adjectif multiples lors de la production des syntagmes composants. Par exemple, le GCP "fleur petite bleue" peut être décomposé en "fleur bleue" ou "fleur petite" pour ne pas perdre trop d'information originelle.

Exemple de décomposition d'un GCP

Soit le GCP "limite des équipements des écoles", pour lequel le pattern-matching a échoué. Dans une première cassure, ce GCP est transformé en: "limite des équipements" et "équipements des écoles".
Si le pattern-matching réussit avec ces deux GCP, la procédure de cassure est terminée et les deux GCP sont considérés comme GCF.
Dans le cas contraire, la procédure de cassure est relancée pour chacun des GCP pour lesquels aucune clique n'a été trouvée, et ainsi de suite. Elle se termine lorsque le pattern-matching réussit pour toutes les sous-structures générées ou lorsqu'il n'y a plus de cassure possible.

Il est nécessaire de s'assurer, qu'au cours du traitement parallèle de plusieurs GCP, il n'y a pas de redondance d'information; autrement dit, que plusieurs décompositions distinctes n'aboutissent pas au même GCP.

3.1. Stratégie de cassure d'un GCP

La cassure syntaxique d'un GCP est fonction de sa catégorie (MI, GCPS, GCPC):

a- Cassure d'un mot isolé:
Elle est forcément impossible et le mot isolé est rejeté.

b- Cassure syntaxique d'un GCPS:
Pour cette catégorie de groupes, on filtre les décompositions possibles selon des critères de consistence de l'information dérivée. Pour les modèles syntaxiques 'SS', 'SPS'(sans article), cette opération n'est pas envisageable car, comme il l'a été souligné en II-2, les syntagmes nominaux de ce type sont indécomposables sans en altérer profondément le

sens.

Pour le modèle 'SPVINF', cette cassure est tolérée dans le cas où l'infinitif est un verbe outil, et est donc fréquent et sans importance (EX: document à voir).
A cet effet, on peut envisager l'exploitation d'une liste, établie a priori après une étude sur la classification des verbes, et qui regrouperait les verbes outils du domaine couvert. Cette liste pourrait être utilisée lors de la cassure des GCPS de ce type: si l'infinitif est inclus dans la liste des verbes outils, le GCPS se réduit à la seule sous-structure possible S; dans le cas contraire, le GCPS est rejeté.

Le modèle 'SA' dérive la seule sous-structure valide 'S'.

Le modèle 'SPS' (cf IV-1-1) dérive également la même sous-structure valide 'S', qui donne deux mots isolés correspondant aux substantifs.

c- Cassure syntaxique d'un GCPC:
A ce niveau, on considère les deux manières de formation possibles du GCPC.

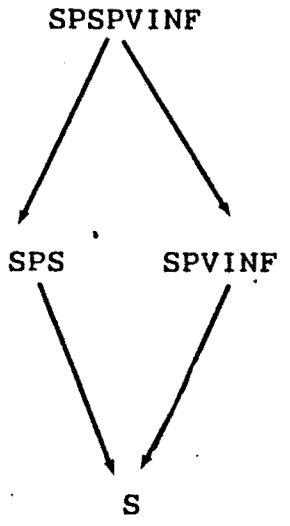
- Le GCPC est le résultat d'une connection de deux GCP, par le biais d'une relation binaire extraite d'un groupe verbal.
  Dans ce cas, la cassure s'effectue au niveau de cette relation, et redonne donc les deux GCPS initiaux.
- Le GCPC a été extrait tel quel du texte:
  Pour sa cassure, la génération du réseau dynamique et essentiellement l'exploitation du réseau statique sont nécessaires.

3.2. Le réseau syntaxique statique

3.2.1. Définition:

Ce réseau regroupe l'ensemble des modèles syntaxiques des GCP. Pour chaque modèle, il permet de décrire l'ensemble des sous-structures valides qui en découlent.
Son organisation est nécessairement hiérarchique, pour répondre à l'objectif visé dans la cassure syntaxique.
Une entrée du réseau correspond à un modèle syntaxique et pointe vers les deux sous-structures maximales de ce modèle et ceci d'une manière récurrente.

EXEMPLE
La décomposition du modèle 'SPSPVINF' donne:



REMARQUE: N structures syntaxiques distinctes peuvent engendrer un ou plusieurs sous modèles communs; dans l'exemple ci-dessus, 'S' est le modèle de dérivation commun à 'SPS' et 'SPVINF'.

On a exclu de la définition du réseau les adjectifs, qui figurent pourtant dans la définition des GCP; cette

simplification (qui n'implique pas que les adjectifs soient ignorés) correspond à deux objectifs:

- Limiter le nombre d'entrées du réseau; pour le simple modèle 'SAPSA', par exemple,, on obtiendrait cinq sous-structures valides: 'SAPS', 'SPSA', 'SPS', 'SA', 'S'.
  En ignorant les adjectifs, la seule sous-structure possible est 'S'; le temps de traitement étant lié au nombre d'entrées du réseau, il y a là un gain de performances substantiel.

- Le retrait temporaire d'un adjectif dans une structure n'entraine pas l'invalidation syntaxique d'un GCP (mots facultatifs). D'où l'inutilité de se référer à un réseau syntaxique complet; il suffit de ne conserver que la règle d'association substantif-adjectif, et de conserver ses instanciations dans un réseau dynamique (cf IV.3.2).

3.2.2. Utilisation du réseau statique:

Le réseau statique permet d'engendrer la ou les deux sous-structures maximales pour un GCPC en cours de traitement. Cette génération est récursive, jusqu'à ce que le pattern-matching réussisse ou devienne impossible.

Dans la cassure d'un GCP il est nécessaire de respecter l'ordre initial des mots pour ne pas introduire d'ambiguité. Ceci est en fait un problème d'autorisation d'instanciation entre une sous-structure syntaxique, extraite du réseau, et les mots qui forment le GCP initial.

Exemple d'instanciation incorrecte:

La structure syntaxique du GCP "limite des équipements des écoles" est 'SPSPS'. La seule sous structure maximale de ce modèle, extraite du réseau, est 'SPS'.
Si on considère l'ensemble des instanciations possibles de

'SPS' avec les éléments du GCP, on obtient:

-"limite des équipements".

-"équipements des écoles".

-"limite des écoles".

Il est évident que le sens de la troisième décomposition n'est pas véhiculé par le GCP initial; cette décomposition ne doit donc pas être retenue.

Pour éviter ce type de décomposition incompatible avec le sens du GCP initial, l'instanciation est réalisée uniquement à partir de mots consécutifs du GCP initial: dans l'exemple précédent, la dérivation "limite des écoles" est ainsi évitée.

3.3. Le réseau dynamique

Il a pour but d'extraire le plus long terme d'indexation possible d'un GCPC initial contenant des adjectifs, et pour lequel le pattern-matching a échoué.
Cette extraction est fondée sur la considération séparée de l'ensemble des adjectifs entrant dans la composition du groupe; elle se déroule selon le schéma suivant:

1. Epuration du groupe initial de l'ensemble de ses adjectifs qui forment alors une liste indépendante.
   Le résultat de cette opération donne une trame du groupe initial.

2. Pattern-matching entre la trame ainsi obtenue, et le thésaurus. Deux cas peuvent alors se produire:

   2.1 Le pattern-matching de la trame réussit:
   Le but recherché est alors la formation du plus long terme d'indexation possible à partir de la trame et de la liste d'adjectifs, en en réintroduisant un maximum (tant

que le groupe ainsi enrichi continue de correspondre à la clique).
Cette phase peut être très coûteuse (algorithme de l'ordre de 2**N, voir en VII réalisation).
Dans le cas où, aucun adjectif ne peut être réintroduit le GCF retenu est la trame du GCPC initial.

Exemple: soit le groupe initial "S1 A1 P1 S2 A2 P2 S3 A3", (les chiffres expriment des instanciations différentes des catégories 'S', 'P' et 'A') pour lequel le pattern-matching a échoué.

- La trame de ce groupe est "S1 P1 S2 P2 S3".

- La liste d'adjectifs qui lui est associée est: ("A1", "A2", "A3").

  Le pattern-matching entre "S1 P1 S2 P2 S3" et le thésaurus réussit.

  On réintroduit un à un les adjectifs de la liste, et le processus continue tant que cette réinsertion demeure compatible avec la clique de référence.

- Parmi tous les sous-ensembles d'adjectifs qui correspondent à la clique en cours de traitement, on retient la première combinaison de cardinal maximum obtenue. Le groupement d'indexation est alors la trame de départ enrichie de cet ensemble d'adjectifs.

- Pour des raisons de coût, on ne retient pas les autres combinaisons maximales possibles (mais cela pourrait être facilement réalisé).

2.2 Le pattern-matching de la trame ne réussit pas:
Dans ce cas l'exploitation du réseau statique s'impose pour décomposer le groupe en GCPS composants.
Si cette opération conduit à des sous-groupes contenant eux même des adjectifs, le même procédé (extraction de la trame, utilisation du réseau dynamique) leur est indépendamment appliqué.

Pendant cette phase de construction des termes d'indexation d'une entité textuelle, on associe à chaque groupe d'indexation sa fréquence d'occurrence totale dans le corpus et sa fréquence d'occurrence locale dans l'unité traitée.
Ces deux informations permettront de valuer la relation d'indexation entre le groupe d'indexation et cette entité textuelle; ceci fait l'objet du chapitre suivant.

4. INDEXATION IMPLICITE

La stratégie décrite ci-dessus consiste donc à extraire le plus long GCF possible d'un GCP trouvé dans le texte. Dans le cas général, ce terme d'indexation pourrait à son tour être décomposé en sous-structures représentant elles mêmes des concepts moins précis. On n'effectue évidemment pas la décomposition pour des raisons de temps d'exécution et de place de stockage des termes d'indexation.
On peut cependant remarquer que par construction, tout GCF indexe implicitement un ensemble d'entités par le biais de ses décompositions possibles.
Si on reprend l'exemple précédent et qu'on considère qu'un ensemble d'unités a été indexé par le terme "équipement des écoles", les concepts sous-jacents de plus haut niveau sont repérés par "équipement" et "école". Ce même ensemble d'unités est implicitement indexé par les termes "école" et "équipement". Retrouver l'ensemble des unités indexées implicitement par "école" revient à considérer l'ensemble des GCF contenant ce mot

(par exemple "équipement des écoles", "école primaire", ...); la réponse est constituée par l'union des ensembles d'unités indexées par ces GCF englobants.
Naturellement, cette indexation implicite correspond à un degré de précision moindre; elle peut néanmoins être utilisée, par exemple, lors du traitement des requêtes pour avoir accès aux documents via des critères plus larges que les termes d'indexation retenus initialement.

CHAPITRE V

CRITERES STATISTIQUES

1. PRINCIPES GENERAUX

Les deux étapes précédentes (analyse syntaxique et utilisation du thésaurus) ont constitué la part majeure du traitement d'indexation de toute unité feuille dans la structure du texte d'entrée:

- L'exploitation des propriétés syntaxiques a permis de déterminer l'ensemble des GCP qui constituent notre représentation du contenu sémantique de l'unité d'entrée.

- Le pattern-matching entre ces GCP et le contenu du thésaurus a permis d'une part le rejet ou la fragmentation de certains GCP, et d'autre part la normalisation et la traduction de ces GCP par des éléments du thésaurus appelés GCF.

Il reste à présent à valuer la relation d'indexation entre les termes d'indexation obtenus et l'unité textuelle dont ils sont issus.

Nous proposons, plutôt qu'une stratégie de sélection, une pondération dans l'intervalle [0, 1] des éléments de la relation d'indexation. Cette option nous paraît plus intéressante qu'une sélection car elle permet:

- Une mise à jour assez aisée de la relation d'indexation dans le cas où le corpus évolue. Dans nos critères statistiques, l'importance d'un terme d'indexation est fonction du comportement de ce terme dans la totalité du corpus; ce comportement étant sensible à l'évolution du corpus,

l'indexation par certains termes peut être remise en cause, ce qui est très difficile si des termes ont été préalablement sélectionnés; il se peut alors que des termes ayant été rejetés deviennent de bons termes d'indexation, mais comment les réintroduire? Il faudrait alors reprendre complètement l'indexation, ce qui serait évidemment très coûteux.
Dans le cas d'une relation pondérée, il y a seulement réévalution du poids associé aux termes déjà existants trouvés dans les éléments nouveaux du corpus et naturellement création éventuelle de termes nouveaux.

- d'échapper au problème du choix du seuil de sélection; la définition d'un seuil mathématique, pour le rejet de certains termes, est délicate et souvent trop brutale, ce qui est une source d'incohérence dans le processus d'indexation.
  Dans les systèmes fonctionnant ainsi, ce seuil (heuristique) est par ailleurs souvent remis en cause.

- de hiérarchiser les réponses à une requête: cette hiérarchie peut être établie en fonction des poids associés dans le thésaurus aux concepts trouvés dans la requête, et de sélectionner dans l'ensemble des entités textuelles celles qui leur sont le plus fortement associées.

La pondération de la relation d'indexation proposée ci-dessous est basée sur deux observations: le comportement du GCF (auquel est associé le poids) au niveau global du corpus et le comportement du même GCF au niveau local dans l'unité traitée.

Compte tenu de l'existence de cette pondération, on peut donner la définition suivante de la relation d'indexation (cf II):
c'est une relation ternaire dont le premier argument est l'ensemble UI des unités d'indexation, le second l'ensemble GCF des termes d'indexation, et le troisième l'ensemble des réels compris entre 0 et 1.

$RI \subset |UI| \times |GCF| \times [0, 1]$

2. PONDERATION DE LA RELATION D'INDEXATION

Comme c'est le cas dans la majorité des méthodes automatiques [AND 73A], [AND 73B], [BOO 75], [COO 78], [FLU 77], [MAR 77], [SAL 71], [SAL 75], [SAL 83], [SPA 71], [WUS 83], ces poids sont calculés selon des méthodes statistiques basées sur la fréquence d'occurrence des GCF dans les diverses unités d'indexation du corpus.
En ce qui nous concerne, nous fondons cette mesure sur une évaluation statistique des notions de 'représentativité d'un concept dans un document' et de 'représentativité d'un document par rapport à un concept' (cf I). Ces notions sont détaillées dans la suite du chapitre.

2.1. Définitions préalables

On note:
- $D=|d|$: l'ensemble des unités d'indexation du corpus D, et ND son cardinal.
- $T=|t|$: l'ensemble des groupes conceptuels finaux (GCF) et NT son cardinal.
- TI(d): ensemble des termes d'indexation d'une unité d. $TI(d) < T$.

A chaque élément de T on associe les informations suivantes:
- FDOC(t) : nombre d'unités dans D ou t existe ;
- FTOT(t): nombre d'occurrences total de t dans le corpus ;
- FLOC(t,d) : nombre d'occurrence de t dans l'unité d'indexation d;

On a aussi:
- TAILL(d): taille d'une unité d'indexation, considérée comme

une suite de termes d'indexation $TAILL(d) = \sum_{t \in T(d)} FLOC(t, d)$.

Pour tout groupe d'indexation, que l'on notera t par la suite, les quantités FDOC(t), FTOT(t), FLOC(t,d) ont été déterminées lors de la phase de pré-indexation, plus exactement au niveau du pattern-matching avec le thésaurus.

Pour l'évaluation du poids d'indexation de t, on tient compte des deux critères suivants:

2.2. Représentativité mutuelle entre un terme et une unité d'indexation

Au niveau local, il est intéressant de reprendre les notions suivantes, déjà présentées au chapitre I:

- La représentativité de t par rapport à d, notée REP(t,d), qui exprime la mesure dans laquelle t participe à l'expression de l'information véhiculée par d, par rapport aux autres concepts exprimés dans d.

- La représentativité de d par rapport à t, notée REP(d,t), qui définit inversement dans quelle mesure d concourt à l'expression de t, par rapport à l'ensemble des autres unités d'indexation qui y font référence.

Examinons plus en détails la définition de ces deux critères locaux.

2.2.1. Représentativité de d par rapport à t

REP(d,t) est relative au comportement du terme t dans le sous-ensemble d'éléments de D qui le contiennent.
Elle mesure la "concentration" de t dans d, par rapport à ce sous-ensemble.

REP(d,t) est fonction du nombre d'occurrences de t dans le corpus, soit FTOT(t), et du nombre d'occurrences de t dans l'unité d'indexation d, soit FLOC(t,d):

$$REP(d,t) = \frac{FLOC(t,d)}{FTOT(t)}$$

ce qui est équivallent à:

$$REP(d,t) = \frac{FTOT(t) - A(t,j)}{FTOT(t)}$$

où $A(t,j) = \sum_{j \neq d} FLOC(t,j)$.

A(t,j) est en fait le complémentaire de FLOC(t,d) par rapport à FTOT(t)et sera noté $\overline{FLOC}(t,d)^f$; dans ce cas l'expression REP(d,t) est égale à:

$$REP(d,t) = 1 - \frac{\overline{FLOC}(t,d)^f}{FTOT(t)} \in [0, 1]$$

En particulier:

- REP(d,t) est donc voisin de 1 lorsque t n'est essentiellement présent que dans d; le concept t est alors essentiellement défini par d, et t discrimine d.
- REP(d,t) est voisin de 0 lorsque t est davantage présent dans le complémentaire de d que dans d; le concept t n'est donc que très faiblement décrit par d.

2.2.2. <u>Représentativité de t par rapport à d</u>

REP(t,d) est relative au comportement de t dans d: elle mesure le taux d'information véhiculée par t, par rapport à l'information totale véhiculée par les termes de d.

REP(t,d) est donc définie comme étant fonction de la quantité FLOC(t,d) et de la taille totale de d, exprimée comme le cumul des fréquences locales de tous les concepts existants dans d.
On a:

$$REP(t,d) = \frac{FLOCT(t,d)}{TAILL(d)}$$

où $TAILL(d) = \sum_{j \in d} FLOC(j,d)$.

ou encore:

$$REP(t,d) = \frac{TAILL(d) - A(t,d)}{TAILL(d)}$$

où $A(t,d) = \sum_{i \neq t} FLOC(i,d)$.

La quantité A(t,d) est le complémentaire de FLOC(t,d) par rapport à TAILL(d), et sera notée $\overline{FLOC}(t,d)^{I}$.
On aura donc, l'expression suivante:

$$REP(t,d) = 1 - \frac{\overline{FLOC}(t,d)^{I}}{TAILL(d)} \in [0, 1]$$

Les valeurs limites de REP(t,d) seront interprétées de la façon suivante:

- REP(t,d) voisin de 1 : t est le sujet primordial de d.
- REP(t,d) voisin de 0 : t est insignifiant dans d.

2.2.3. Représentativité mutuelle entre t et d

On définit la valuation, notée F(t,d), de la relation d'indexation, concernant un terme t et une unité textuelle d par:

$$F(t,d) = \frac{REP(t,d) + REP(d,t)}{2} \in [0, 1]$$

F(t,d) indique dans quelle mesure l'existence de t dans d, discrimine une information 'consistante' par rapport au reste du corpus.
On peut avoir les quatre cas limites suivants:

C1-REP(t,d) et REP(d,t) sont voisins de 1: t est décrit presque exclusivement par d, et le contenu de d est très fortement exprimé par t; t est un bon descripteur de d et le discrimine dans le corpus. L'unité d constitue la meilleure réponse pour t.

C2-REP(t,d) et REP(d,t) sont tous deux voisins de 0: la présence de t dans d est insignifiante et d concourt faiblement à la définition de ce concept; t est donc mauvais descripteur pour d. L'unité d ne constitue pas une réponse satisfaisante pour t.

C3-REP(t,d) est voisin de 1 mais REP(d,t) est insignifiant: t exprime fortement le contenu de d mais il est présent dans plusieurs autres documents du corpus; t décrit bien le contenu de d mais n'a pas un bon pouvoir discriminant. L'unité d constitue donc une bonne réponse pour t qui est un terme général; sauf critère de sélection supplémentaire dans la requête, il faut reformuler celle-ci avec un terme spécifique de t.

C4-REP(t,d) est insignifiant mais REP(d,t) est voisin de 1: dans ce cas t est entièrement défini par d mais n'est pas fortement impliqué dans son contenu. L'unité d est la meilleure réponse possible pour t dans le corpus mais ce terme exprime une notion très marginale par rapport au thème général. Il faut reformuler la requête dans le sens d'un élargissement vers des termes génériques de t.

Remarque.
A la fin de la valuation de la relation d'indexation entre un GCF et une unité donnée, la fréquence d'occurrence de ce GCF est répercutée sur la fréquence totale respective à l'ensemble des GCF composants et existants déjà dans la base. Cette répercussion permet d'augmenter la fréquence totale associée à un GCF composant, et par conséquent de diminuer les poids d'indexation qui lui sont associés. Cette opération permet donc de banaliser les concepts génériques en tant que termes d'indexation, et par conséquent de favoriser les concepts plus précis.

En conclusion, la mesure proposée donne une indication globale quant à la représentativité mutuelle entre un terme et une entité de structure. Dans les cas extrêmes, cette mesure permet de s'orienter, à l'interrogation, vers l'acceptation ou le rejet d'une entité de structure du corpus lors de l'élaboration de la réponse. Pour les valeurs intermédiaires, il y a nécessité d'une reformulation de la requête vers des termes plus généraux ou au contraire vers des termes plus spécifiques. Cette décision peut être prise en considérant alors les valeurs respectives de REP(t,d) et REP(d,t).

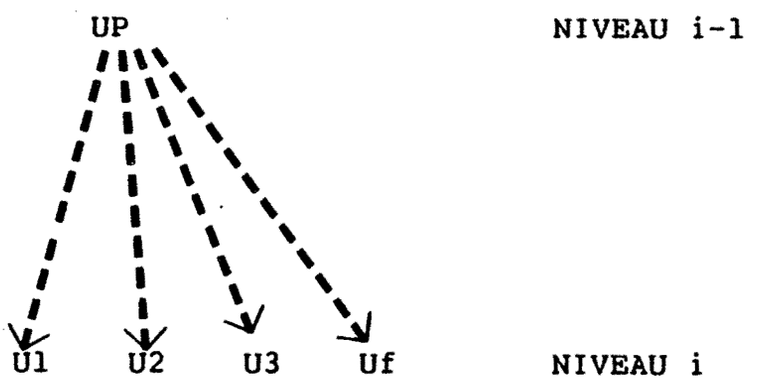
CHAPITRE VI

INDEXATION DYNAMIQUE

1. PRESENTATION

La stratégie d'indexation dynamique repose sur l'aspect arborescent des documents et sur des indicateurs tels que les titres des entités de cette structure; elle concerne un processus ascendant d'indexation depuis les unités d'indexation minimales, jusqu'aux unités d'indexation maximales (cf II).

Si on considère, par exemple, la structure suivante:



où l'entité de structure UP, de niveau (i-1), est composée de f entités filles, le processus d'indexation se déroule en deux temps:

- U1,...,Uf sont analysées et indexées de manière indépendante, selon le processus vu dans les deux chapitres précédents.

- UP sera éventuellement indexée par un sous-ensemble de termes d'indexation de ses entités filles U1, U2,...,Uf.

- Le processus reprend en indexant le niveau i-2 à partir des entités filles de niveau i-1 que l'on vient de traiter.

Il est important de noter que cette remontée ne nécessite aucune analyse complémentaire du texte, si ce n'est la prise en compte des titres des entités de niveaux supérieurs.

2. ANALYSE D'UNE UNITE D'INDEXATION MINIMALE

Dans chaque feuille de la structure du document peuvent coexister divers formalismes (langue naturelle, programme, schéma...).

Un élément en langue naturelle est donc transformé en un ensemble de GCF du thésaurus pour lesquels la relation d'indexation est valuée de la manière décrite en V.
Nous avons vu (cf IV.2) que tout titre informatif est contenu dans une clique; à l'indexation, le GCF correspondant aura une valuation forcée à 1 (valeur maximale) avec l'unité contenant ce titre.

Il se peut que l'entité de niveau minimal contienne des sections spéciales de type INTRODUCTION, CONCLUSION,... Si ces éléments ne participent pas à une structuration physique de l'entité, on en tient néanmoins compte dans l'indexation en considérant ces titres comme des délimiteurs de portions particulièrement significatives du texte, qui sont traitées implicitement comme décrit en VI.3 ci-dessous.

Dans notre étude, nous ne nous sommes pas encore préoccupé d'outils pour l'analyse d'un formalisme autre que la langue

naturelle (programme, schéma...), bien que ces types d'objets soient disponibles dans la base enregistrée. L'indexation de ces éléments est pour l'instant limitée aux titres ou légendes qui les accompagnent généralement. La stratégie utilisée est alors la même que pour les titres pleins. Ce traitement concerne les programmes, les schémas, les figures et les tableaux, et fait intervenir la table des TITRES-PROCEDES qui renseigne sur le type particulier de l'entité introduite.

3. STRATEGIE DE REMONTEE DES TERMES D'INDEXATION

L'analyse d'une unité d'indexation de niveau quelconque comporte deux aspects:

- Une indexation par le titre plein qui introduit cette unité, et qui est retenu automatiquement comme GCF.

- Une stratégie de remontée des termes d'indexation des unités considérées dans la hiérarchie du document.

La remontée des termes d'indexation d'une unité fille vers le niveau supérieur dans la hiérarchie, est fonction du rôle de cette unité par rapport à l'ensemble. On distingue, de ce point de vue, les cas particuliers de certaines entités (INTRODUCTION, CONCLUSION...), du cas général.

Une unité textuelle ayant pour titre, INTRODUCTION, CONCLUSION, RESUME, GENERALITE, a un aspect résumant commun à l'information véhiculée par l'ensemble des entités de structure ayant même père direct dans la hiérarchie. De même que le titre plein, l'ensemble des termes d'une telle unité est remonté intégralement et avec le même poids d'indexation, vers l'unité directement englobante dans la hiérarchie du document.

La stratégie de remontée relative aux autres types d'unités vers un noeud père, consiste à ne prendre en compte que le sous-ensemble des termes d'indexation communs à toutes les unités filles de ce noeud, et à réévaluer le sous-ensemble de la relation d'indexation pour ces termes, au niveau considéré.

4. VALUATION DU POIDS D'INDEXATION DANS LA STRATEGIE DE REMONTEE

Lors de la remontée d'un terme vers une unité père de la structure, la valuation de la relation d'indexation est effectuée selon un principe analogue à celui défini précédemment: on réévalue la représentativité du terme par rapport à la nouvelle entité de structure, et la représentativité de celle-ci pour le terme par rapport au reste du corpus.

Pour la détermination de ce poids, considérons les définitions suivantes (un sous-ensemble ayant déjà été vu en V):

-di-1 : l'unité père de niveau i-1.

-Di = |dij| : ensemble des Ni unités de niveau i ayant pour père di-1;

-TCi = |tc| : le sous-ensemble de termes d'indexation communs aux éléments de Di;

Le poids d'indexation d'un terme tc par rapport à l'unité di-1 s'exprime donc en fonction de la représentativité de tc par rapport à di-1 (REP(tc,di-1)) et de la représentativité de di-1 par rapport à tc (REP(di-1,tc)). Ce poids, noté F(tc,di-1), est défini comme en V par:

$$F(tc,di-1) = \frac{REP(tc,di-1) + REP(di-1,tc)}{2}$$

a) <u>Calcul de la représentativité de tc par rapport a l'unité di-1</u>

Cette quantité est fonction des deux éléments FLOC(tc,di-1) et TAILL(di-1); selon la formule donnée en V, elle est égale à:

$$REP(tc,di-1) = \frac{FLOC(tc,di-1)}{TAILL(di-1)}$$

Le terme tc étant commun aux éléments de l'ensemble Di, les quantités FLOC(tc,di-1) et TAILL(di-1) peuvent être toutes deux exprimées par les informations relatives au terme tc dans les éléments de Di:

- La fréquence locale de tc dans di-1 est la somme des fréquences locales de tc dans les éléments de Di, d'où:

  A) $FLOC(tc,di-1) = \sum_j FLOC(tc,dij)$.

- La taille de l'unité di-1 n'est rien d'autre que la somme des tailles des éléments de Di; on a donc:

  B) $TAILL(di-1) = \sum_j TAILL(dij)$.

En utilisant les définitions A et B dans l'expression de REP(tc,di-1) on obtient:

$$REP(tc,di-1) = \frac{\sum_j FLOC(tc,dij)}{\sum_j TAILL(dij)}$$

En introduisant, dans chaque quantité FLOC(tc,dij) la quantité TAILL(dij) on obtient:

$$REP(tc,di-1)=\frac{\sum_j (FLOC(tc,dij)/TAILL(dij))*TAILL(dij)}{\sum_j TAILL(dij)}$$

L'expression 'FLOC(tc,dij)/TAILL(dij)', pour j donné, n'est autre que la représentativité de tc par rapport à l'unité dij, telle qu'elle a été calculée au niveau i.

L'expression 'TAILL(dij)/$\sum_K$ TAILL(dik)' représente la fraction d'information véhiculée par l'unité dij par rapport à l'entité père di-1. Elle sera notée FR(dij,di-1).

Dans ce cas la définition de REP(tc,di-1) devient:

$$REP(tc,di-1)=\sum_j REP(tc,dij) * FR(dij,di-1)$$

De cette formule, il ressort que la représentativité de tc par rapport à di-1 est fonction de la représentativité de tc par rapport à chacune des unités filles de di-1, et de la taille relative de ces unités par rapport à di-1.

b) <u>Calcul de la représentativité de di-1 par rapport à tc</u>

Cette quantité est fonction de la fréquence locale de tc dans di-1, FLOC(tc,di-1), et de la fréquence d'occurrence de tc dans le corpus, FTOT(tc). Comme en V, elle est définie par:

$$REP(di-1,tc)=\frac{FLOC(tc,di-1)}{FTOT(tc)}$$

On peut aussi écrire d'après (A):

$$REP(di-1,tc) = \frac{\sum_j FLOC(tc,dij)}{FTOT(tc)}$$

La quantité 'FLOC(tc,dij)/FTOT(tc)', pour j donné, est en fait la représentativité de l'unité dij par rapport à tc 'REP(dij,tc)'; d'où:

$$REP(di-1,tc) = \sum_j REP(dij,tc);$$

La représentativité de di-1 par rapport à tc est donc égale à la somme des représentativités de chacune de ses unités filles par rapport à tc.

c) <u>Variation du poids d'un terme d'indexation à la remontée</u>

On s'intéresse ici à l'évolution des notions de représentativité d'un terme lors de l'indexation des niveaux de structure supérieurs. Les modalités de pondération de la relation d'indexation décrites ci-dessus tiennent compte du niveau de structure considéré; on peut en particulier en déduire, pour un terme donné, quel est le niveau pour lequel les critères de représentativité sont les plus élevé
Analysons les variations de la mesure en fonction du niveau d'unité considéré:

- REP(di-1,tc) est supérieure à REP(dij,tc) quel que soit j.
- REP(tc,di-1) est fonction de REP(tc,dij), pour j donné et de la fraction d'information véhiculée par l'unité fille dij.

Une condition suffisante pour que F(tc,di-1) soit supérieur à F(tc,dij), pour j donné, est que REP(tc,di-1) soit supérieure ou égale à REP(tc,dij); ceci se traduit ainsi:

$$\frac{FLOC(tc,dij)}{TAILL(dij)} \leqslant \frac{FLOC(tc,di-1)}{TAILL(di-1)}$$

où:

$$FLOC(tc,dij) \leqslant FLOC(tc,di-1) \times \frac{TAILL(dij)}{TAILL(di-1)}$$

ce qui donne:

$$FLOC(tc,dij) \leqslant FLOC(tc,di-1) \times FR(dij,di-1)$$

En particulier, lorsque les tailles respectives aux filles de di-1 sont toutes égales, cette inéquation devient:

$$FLOC(tc,dij) \leqslant \frac{FLOC(tc,di-1)}{NI}$$

Dans ce cas, la quantité de droite représente la fréquence locale moyenne de tc, par rapport à l'ensemble des filles du noeud père di-1.

Pour un terme donné tc, la représentativité à un niveau plus élevé dans la structure peut donc être supérieure ou inférieure à celle obtenue au niveau considéré. Le sens de cette variation dépend du contenu des autres unités de même niveau: si le concept correspondant est davantage précisé au niveau supérieur, il y aura augmentation de la mesure, sinon il y aura diminution (le concept est très estompé au niveau supérieur car les entités de même niveau de départ concernent d'autres sujets).

5. CONCLUSION

Il ressort de ces définitions que la remontée des termes d'indexation vers les entités pères s'effectue très facilement par sélection des GCF communs, puis par un calcul récurrent très

simple appliqué à ces groupes d'indexation.
Un terme donné indexant une entité de niveau i, peut être considéré comme terme d'indexation de l'entité père de niveau i-1 si et seulement si ce terme est commun à l'ensemble des entités filles correspondant à ce sous-arbre. Il est alors possible que la relation d'indexation avec l'entité père corresponde à un poids plus élevé qu'avec l'une quelconque des entités filles. Cela signifie que le niveau père est plus compatible avec le degré de généralité du terme, que les entités dépendantes. Il est également possible que le phénomène inverse se produise, auquel cas les entités filles sont considérées comme meilleures réponses, pour le terme considéré, que l'entité père où le concept associé apparaît comme beaucoup plus marginal. La stratégie proposée permet donc une adaptation de la réponse au niveau structurel le plus adéquat du texte.

CHAPITRE VII

REALISATIONS

1. INTRODUCTION

Les réalisations pour l'application ont porté sur le développement de trois outils principaux qui à partir d'un texte initial réalisent son indexation complète.

1) Le module STRUCTDOC qui permet l'acquisition d'un texte et son enregistrement sous forme de structure arborescente (cf VII.2).

2) Le module de traitement linguistique, composé des modules LEMMAT et ANAL (cf VII.3).
   Le module LEMMAT permet la lemmatisation du vocabulaire initial du texte lu: ce module remplace chaque terme de ce texte par une forme normalisée (les différentes formes grammaticales d'un terme sont ramenées à une forme canonique à laquelle est associée une catégorie grammaticale).
   Le module ANAL permet de reconnaître et de former à partir du texte lemmatisé les éléments ayant une structure syntaxique de GCP.

3) Le module d'indexation PROCINDEX, composé de quatre fonctions principales (cf VII.4), permet d'établir et de valuer la relation d'indexation entre un texte analysé via les modules linguistiques, et l'ensemble des éléments d'indexation qu'on peut en extraire. Ceci est effectué par référence au thésaurus dont le contenu, constitué des concepts du domaine, est décrit en VIII.

L'enchaînement de ces trois outils lors de l'analyse d'un texte est identique à l'ordre de leur présentation dans ce chapitre.

Une présentation des algorithmes essentiels est donnée pour chacune de ces fonctions, de façon à avoir un exposé aussi condensé et homogène que possible de programmes ayant été par ailleurs réalisés à l'aide de différents langages.

L'ensemble des programmes correspondants a été développé sur un HB68 sous le système MULTICS dont la capacité mémoire très importante et la puissance de calcul sont indispensables à notre type d'application mettant en oeuvre des masses d'informations importantes et des traitements complexes.
Pour le moment ces programmes ont un caractère expérimental; leur intérêt immédiat est principalement la validation qualitative de la stratégie proposée.

2. DESCRIPTION ET STOCKAGE DU TEXTE PAR STRUCTDOC

Ce module produit la structure arborescente d'un texte et est paramétré par le type de documents traités dans le corpus. Ceci permet de définir a priori les niveaux d'unité d'indexation minimale et maximale choisis pour le type de document (cf II). Le texte d'entrée est saisi selon des modalités particulières permettant le décodage de sa structure logique. Sans entrer ici dans les détails de ce formalisme, on peut considérer qu'il est composé essentiellement d'opérateurs d'indentation de structure et de délimiteurs d'objets textuels particuliers (titres...).

Un texte d'entrée est donc analysé et transformé en deux éléments:
- ARBRDOC qui donne la description arborescente du texte.
- INFODOC qui contient l'information textuelle proprement dite.

Chaque entrée de ARBRDOC décrit une entité de structure du texte et pointe dans INFODOC sur le contenu de cette entité. Pour les noeuds non terminaux de la structure, la seule information

textuelle associée est le titre; pour les feuilles de l'arborescence, on trouve également le contenu textuel complet de l'entité.
Dans ARBRDOC, à chaque entité est associé d'une part l'ensemble des pointeurs vers la description des entités filles directes dans la hiérarchie, et d'autre part un pointeur vers la description du noeud père de cette entité dans la hiérarchie.

Dans tous les autres programmes développés pour l'application, il est nécessaire de passer par la structure arborescente pour accèder au contenu textuel d'une entité.

Donc globalement STRUCTDOC reconnaît une entité de structure dans le texte d'entrée, produit sa description arborescente dans ARBRDOC et stocke les informations textuelles dans INFODOC.

Le programme correspondant est composé de 6 modules qui sont tous codés en PASCAL:

a- Traitement des documents: reconnaît le type du document à partir des paramètres d'entrée et appelle le module de traitement correspondant.

b- Traitement d'un type particulier de document: module de lancement pour un document de type notice technique (cf VIII) qui correspond au corpus traité dans notre expérimentation.

c- Structuration d'un type de document: module récursif pour le traitement d'une entité de structure donnée et de ses filles directes. Ce module reconnaît le début d'une entité de structure dans le texte; il délimite le texte qui lui correspond et décrit sa structure arborescente dans ARBRDOC. Il appelle les deux modules suivants pour le stockage du texte associé à l'entité de structure. Ce sont ces deux derniers modules qui produisent le fichier INFODOC des

données textuelles.

d- Traitement des titres: reconnaît le titre de l'entité de structure et le délimite par des symboles spéciaux. Il appelle le module ci-dessous pour le traitement du texte qui suit le titre.

e- Stockage du texte: à partir de la ligne lue dans le texte d'entrée, ce module reconnaît les caractères d'indentation inclus dans le texte. Il les supprime et recopie la ligne ainsi réduite dans INFODOC.

f- Traitement de texte 'plat': en dessous du niveau correspondant à l'unité d'indexation minimale, on ne tient plus compte de l'arborescence du texte.
Ce module traite donc le texte correspondant de manière non structurée (mais les titres qu'il contient sont toujours délimités de façon à pouvoir les réutiliser lors de l'indexation).

3. OUTILS LINGUISTIQUES

L'outil linguistique se compose de deux modules: le module LEMMAT de lemmatisation du vocabulaire des textes, et le module ANAL qui reconnaît et forme les structures syntaxiquement valides de GCP.

3.1. Lemmatisation du vocabulaire

Cette opération a pour but d'associer à chaque mot nouveau du texte traité, une catégorie grammaticale et un lemme qui est considéré comme la racine morphologique de ce mot nouveau.
Le traitement réalisé au cours de notre expérimentation ne correspond pas à une véritable analyse morphologique des textes

d'entrée, en ce sens que ni le genre ni le nombre ne sont déterminés pour l'élément lu. De plus lorsqu'il y a ambiguité sur des éléments homographes, elle ne peut être levée.
Ce traitement bien qu'incomplet par rapport à une analyse morphologique normale, est toutefois indispensable pour assurer une normalisation du vocabulaire et obtenir les structures de GCP recherchées dans le texte (cf VII.3.2).

L'analyse morphologique est semi-automatique et considère deux lexiques: LEXIQUE qui regroupe le vocabulaire initial des textes lus, et LEXIQUELEM dans lequel, pour chaque élément de LEXIQUE, sont fournis le lemme et la catégorie grammaticale correspondants.
Malgré toutes ses insuffisances actuelles, cette opération a néanmmoins pu être effectuée dans d'assez bonnes conditions étant donné que le texte traité (notice technique) ne présentait pas de grandes difficultés d'analyse (vocabulaire restreint, syntaxe simple). Un logiciel beaucoup plus élaboré est en cours de définition dans le groupe et doit être incorporé dans la chaine d'analyse des documents [PAL 81].

3,2. L'analyse syntaxique de surface via ANAL

Une phrase étant lemmatisée (tous les mots sont représentés par leur lemme respectif via le module de lemmatisation), elle est ensuite transformée par l'analyseur ANAL en un ensemble de groupes conceptuels primaires (GCP).

Pour des raisons de coût, la syntaxe de GCP considérée à titre expérimental dans ANAL a été limitée par rapport à celle donnée en IV-1.
Actuellement tout GCP est formé à partir d'un seul groupe nominal du texte analysé, et les verbes de la phrase ne sont pas considérés pour leur génération. En particulier, la catégorie nommée GCPC en IV-1 n'est pas entièrement produite.

Rappelons que les GCP sont le résultat d'une transformation de groupes nominaux reconnus dans le texte (cf IV-1) et dont la syntaxe a également été limitée à la suivante:

GP::= G [PREP G];
G::= [ARTICLE] [ADJECTIF]* GS [ADJECTIF]*;
GS::= substantif/ substantif substantif;
PREP::= préposition;

La syntaxe d'un GCP est limitée de la façon suivante:

GCP::= GN [PREP GN];
GN::= GS [ADJECTIF]*;
GS::= substantif/ substantif substantif;
PREP::= préposition;

Cette limitation a été nécessaire dans le but d'éviter les ambiguités dans la formation des GCP: en effet, le module de lemmatisation fournissant une morphologie très pauvre (pour un lemme, on a uniquement sa catégorie grammaticale, le genre et le nombre étant ignorés) il aurait été risqué de vouloir reconnaître une syntaxe plus complète des groupes.

Le traitement d'une phrase se déroule ainsi:

L'analyse d'une phrase lue débute toujours sur un mot de catégorie grammaticale de début possible de groupe nominal (substantif ou adjectif).

En cours d'analyse, si il y a ambiguité sur la formation d'une structure courante de GCP, il y a abandon de cette structure au profit d'une sous structure non ambigue et le reste de la phrase est ignoré jusqu'au prochain début de groupe nominal possible; l'analyse est alors relancée à partir de ce nouveau début.

La formation d'un GCP se fait progressivement depuis une structure très restreinte jusqu'à des structures plus larges, en considérant les mots voisins du mot de début.

Le fonctionnement du module ANAL est fondé sur l'évaluation de trois prédicats principaux:

a- Repérage du début de groupe nominal:
Ce prédicat est évalué à vrai si la catégorie grammaticale du mot courant ne correspond pas à un début de groupe nominal selon la syntaxe (1). Dans ce cas les mots de la phrase sont ignorés. Le prédicat prendra la valeur 'faux' lorsque le mot courant correspond à un début possible de groupe nominal.
Ce prédicat est évalué en début d'analyse et peut être réévalué au cours de la formation d'un GCP si nécessaire.

b- Génération des GCP restreints (GN):
Le prédicat est à vrai si les catégories de la structure courante reconnue correspondent à une structure de groupe nominal où un ou deux substantifs sont éventuellement suivis ou précédés d'adjectifs qualificatifs. Dans ce cas il y a formation d'un GN; si dans le texte un substantif est qualifié à gauche par un ensemble d'adjectifs, dans le GN formé cet ensemble d'ajectifs est placé à droite. Par exemple, à partir de la structure courante 'petite fleur bleue' il sera formé le GN 'fleur petite bleue'. Cette transformation correspond à une normalisation des GCP, de façon à faciliter ultérieurement la composition des termes d'indexation, remarquons qu'elle n'engendre aucune perte d'information.

c- Génération des GCP étendus:
Ce prédicat est évalué à vrai si on a formé une structure valide de GCP . Cette formation débute sur la reconnaissance d'un groupe nominal GN1 suivi éventuellement par une préposition et un groupe nominal GN2 (complément de nom), pour donner une structure complexe correspondant à la règle 1 de la syntaxe des GCP. Dans ce cas l'analyse est relancée à partir de GN2, pour former le groupe suivant.

Exemple: pour le groupe nominal 'limitation de équipement de école' on extrait un premier GCP 'limitation de équipement' et l'analyse est relancée à partir de 'équipement' ce qui donne un second GCP 'équipement de école'.

ANAL est codé en FOLL-PROLOG [DON 83] dérivé de PROLOG et implanté sous forme interprétée sur MULTICS (CICG).
FOLL-PROLOG est encore au stade expérimental et pour des raisons d'efficacité (le temps d'exécution est pour le moment important) et surtout d'homogénéité, ANAL est en cours de redéfinition en LISP (tous les modules assurant l'indexation sont codés en LISP).

4. LE MODULE D'INDEXATION

Pour chaque unité d'indexation traitée, ce module établit et value la relation d'indexation entre cette unité et l'ensemble des GCF qu'on peut extraire des GCP qui y ont été reconnus. Ceci se fait par le biais d'une base de connaissances qui est actuellement structurée (cf IV-2) comme un ensemble de cliques contenant chacune une liste de GCF qui ont été dérivés par le processus de pattern-matching (cf IV-2-3).
Le contenu effectif de la base est décrit en VIII.

Schématiquement chaque unité textuelle traitée est liée via la relation d'indexation à la base de connaissances, de la manière suivante:

BASE DE CONNAISSANCES

UT <---- RI ----> | GCF1, GCF2, ... GCFn |

Le processus d'indexation consiste donc à valuer et à enregistrer dans la clique contenant un GCF donné, le poids

d'indexation entre ce GCF et chacune des unités d'indexation qui le contiennent.

Selon les définitions données en V, la valuation du poids d'indexation entre un GCF et une unité est fonction de la fréquence d'occurrence totale du GCF dans le corpus, la fréquence locale du GCF dans l'unité traitée et la taille en nombre de GCF de cette unité.
L'entrée du module d'indexation est donc un ensemble de GCP obtenu via l'analyse syntaxique de l'unité, et une référence à cette unité dans la structure hiérarchique globale du texte analysé.

On distingue quatre fonctions principales dans le module:

- Pattern-matching: c'est la fonction qui réalise le pattern-matching entre un GCP initial et la base de connaissances. Cette fonction concerne donc toutes les transformations de ce GCP en un ensemble de GCF (cf IV-2-3).

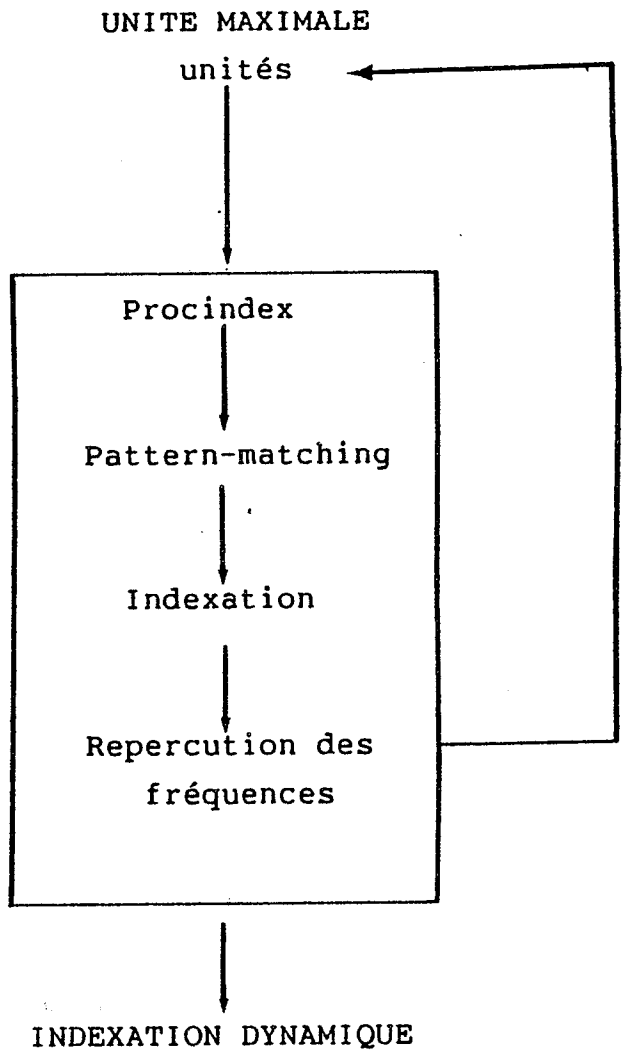
- Indexation: c'est la fonction qui établit et value la relation d'indexation. Elle met à jour la base de connaissances en créant ou en réévaluant la relation d'indexation entre chaque GCF extrait du GCP initial et l'unité d'indexation traitée.

- Répercussion des fréquences: cette fonction met à jour les fréquences totales d'occurrence des GCF de la base. De chaque GCF tiré du GCP initial, cette fonction extrait l'ensemble des GCF plus restreints qui y sont inclus et met à jour leur fréquence totale d'occurrence dans le but de réévaluer les poids d'indexation qui leurs sont associés.

- Remontée des termes d'indexation: cette fonction permet de remonter le sous-ensemble des termes d'indexation commun à

toutes les unités filles d'un noeud et à établir et valuer le sous-ensemble de la relation d'indexation entre ces termes et ce noeud de la structure.
Cette fonction est activée globalement pour une unité d'indexation non minimale et donc à la fin de l'indexation complète des filles de cette unité.

L'enchainement de ces quatre modules est assuré par la fonction PROCINDEX qui lance le processus d'indexation pour une unité d'indexation donnée.
Cet enchainement est le suivant:



Les algorithmes des principales fonctions du module d'indexation sont décrits en annexe I.

L'ensemble des fonctions réalisant PROCINDEX a été codé en MACLISP [MAC 82], dérivé de LISP et dont une version interprétée et une version compilée sont implantées sur MULTICS (CICG).

5. CONCLUSION

A la fin de l'analyse complète d'un texte, à chaque GCF d'une clique est associé l'ensemble des unités de structure qui le contiennent dans le texte.
Une clique initiale (T1, T2, T3, T4) devient une donnée du type suivant:

```
clique(T1 T2 T3 T4)-->(((T1 T2) frectot (frecloci pdi refi)...)
                       ((T2 T4) frectot (freclocj pdj refj)...)
                        ...etc
```

où (('T1 T2', 'T2 T4') sont des GCF de la clique considérée, ('refi', 'refj') sont des références vers les unités traitées et ('pdi', 'pdj') sont les poids correspondants respectivement aux liaisons 'T1 T2' --> 'refi' et 'T2 T4' --> 'refj'.

A tout moment, les poids d'indexation associés à un GCF de la base sont relatifs exactement à l'ensemble de textes traités dans le corpus.

L'implantation de la relation d'indexation en LISP présente deux intérêts majeurs dans notre application:

- Le premier est relatif au stockage de la masse d'informations très importante dans l'application: celle-ci n'aurait pu être mieux compactée vu la gestion de données dans LISP (plusieurs occurrences d'un atome sont représentées par la même donnée pointée).

- Le second est relatif à la mise à jour de la relation d'indexation lors de l'évolution du corpus: cette mise à jour concerne essentiellement l'introduction d'un nouveau texte dans le corpus et par conséquent la réévaluation des divers poids d'indexation d'un GCF extrait de ce texte. Etant donnée la rapidité d'accès aux atomes et à leurs propriétés en LISP (accès direct par HASHCODE) cette mise à jour est rendue très aisée et peu coûteuse.

Néanmoins, comme nous l'avons souligné en introduction, les outils développés pour cette application sont encore expérimentaux, l'intérêt immédiat étant la validation qualitative de la stratégie avancée.
Ceci concerne plus particulièrement les outils linguistiques et l'outil de stockage de document.

Les programmes d'analyse linguistique disponibles actuellement ne permettent pas un véritable traitement de la langue naturelle et parmi nos préoccupations immédiates figure l'intégration d'un analyseur morpho-syntaxique du français (cf IV) plus complet et dont la réalisation est en cours d'achèvement [PAL 81].

L'outil de stockage et d'accès aux documents utilisé dans cette expérimentation n'est en fait que provisoire: notre travail s'insérant dans le cadre du projet CONCERTO [BCK 83], [KBC 83], il sera très prochainement remplacé par MENTOR-RAPPORT dérivé du système MENTOR qui a été développé à l'INRIA [LAN 81], [MEL 83].

Dans le projet CONCERTO, notre travail se situe au niveau de la définition d'une fonction documentation sur un poste de travail. Le choix de ce poste de travail a porté sur une SM90 (multi microprocesseurs sous le système MPX) développé par le CNET et l'INRIA.
Vu la masse importante d'informations dans un corpus, l'application complète ne peut être envisagée sur la SM90 et

nécessite un aspect réparti des traitements:
l'indexation du corpus s'effectuera sur une grosse machine du type HB68 et seuls les résultats de cette analyse seront transférés sur une SM90, où ils seront disponibles pour une fonction interrogation dont le développement va être entrepris. Pour assurer l'homogénéité de l'ensemble, ce module sera développé en LELISP, disponible sur la SM90.

CHAPITRE VIII

EXPERIMENTATION

1. LE CORPUS TRAITE

L'expérimentation en cours a porté sur un corpus de documents de type technique très structurés. Il s'agit des deux tomes des NEF utilisés par le CNET (manuels de normalisation pour les autocommutateurs) [BCK 83], [KBC 83].

Dans la structure de ce type de documents, on a considéré trois niveaux hiérarchiques (chapitre, sous-chapitre, paragraphe et sous-paragraphe).
Le niveau de l'unité d'indexation minimale a été fixé au sous-paragraphe, car dans les NEF ce dernier peut introduire ou définir des concepts qui n'existent pas dans les niveaux supérieurs.
Les chapitres des NEF étant assez indépendants et volumineux, l'unité d'indexation maximale a été fixée au niveau chapitre.

Le premier tome des NEF est structuré en quinze chapitres dont certains assez volumineux en texte en langue naturelle.
Le second tome est structuré en vingts trois chapitres où les programmes, schémas et tableaux représentent un pourcentage très important du texte (ces données, bien qu'existantes dans la base, ne sont pas concernées par l'indexation).

Ces deux documents présentent à eux deux environ 100.000 mots de texte en langue naturelle.
La phase lemmatisation a engendré pour ces 100.000 mots de texte deux lexiques: le premier d'environ 6000 éléments contient le vocabulaire initial des textes traités; le second correspond au vocabulaire lemmatisé associé au lexique initial, et comporte environ 4000 formes.

Les catégories grammaticales principales sont: substantif, nom propre (et sigle technique), adjectif (qualificatif), verbe et préposition. Une répartition du lexique selon les catégories principales a montré la prédominance des catégories substantif, adjectif, avec un pourcentage non négligeable de sigles techniques (80% du vocabulaire lemmatisé total pour ces trois catégories principales).

2. LA BASE DE CONNAISSANCES

L'élaboration de la base s'effectue en deux phases (cf IV):

- Construction de la matrice d'associations des termes du vocabulaire: pour chaque couple de termes du texte traité, la matrice contient une liaison dans l'intervalle [0, 1] qui exprime le degré d'association de ces deux termes.
  La matrice obtenue est symétrique et creuse (pas plus de 4% de taux de remplissage).

- Constitution de la base de connaissance: cela consiste à extraire du graphe de la matrice d'associations l'ensemble des sous-graphes complets maximaux, ou cliques.
  Une clique contient donc un ensemble de termes complètement connectés entre eux et par là même, considérés comme représentatifs d'un concept significatif du corpus (voir exemple ci-dessous).

Les expérimentations pour la constitution de la base de connaissance ont porté sur les deux tomes des NEF [BCK 83], [KBC 83].

A titre d'exemple nous donnons quelques résultats concernant la matrice d'associations, les cliques et leurs traces pseudo-syntaxiques, relatifs à un chapitre important (2853

termes lemmatisés) du tome un des NEF. Le contenu de ce chapitre porte sur la documentation technique.

La matrice d'associations a donné 10658 couples de termes. La liaison n'est différente de 0 que pour 2537 couples.
Pour la construction des cliques, seuls les couples dont la liaison est jugée significative (seuil>= 0,02) ont été considérés.
Le nombre de cliques obtenu est 809, leur cardinal variant entre 1 et 16. Seuls 425 termes sur les 700 termes du vocabulaire du chapitre ont été conservés après cette opération.

La trace pseudo-syntaxique des cliques de ce chapitre montre que, sur un total de 809 cliques, 30,2% d'entre elles correspondent exactement à un modèle syntaxique, toutes les autres incluant au sens strict, un ou plusieurs modèles.
Les trois quarts des cliques contiennent au moins une préposition. D'autre part, on peut remarquer que seulement 33% des cliques contiennent au moins un adjectif, et que le nombre d'adjectifs que l'on peut trouver dans une clique est restreint (parmi cet ensemble de cliques, 79,7% ont un seul adjectif, 19,2% en ont deux et seulement trois cliques (soit 0,01%) possèdent trois adjectifs).

Exemples de cliques obtenus:
(symbole graphique)
(de système télécommunication)
(de ordre lecture)
(de plan sectionnement)
(de service contrôle télécommunication)
(de structure documentation)
(schéma ensemble organe être ou)
(documentation référence microfilm sur)
(d' marché étude)
(document documentation microfilm être sur)

Dans la phase actuelle, les cliques représentent un modèle partiel des connaissances extraites du corpus et peuvent être utilisées à deux niveaux dans un système de recherche d'informations; à l'indexation et à la recherche ultérieure des documents:

a- Le choix de prendre la base des cliques comme base d'indexation dans notre application est délibéré. De part leur construction les cliques représentent en fait une préindexation du corpus: le regroupement de termes dans une clique tient non seulement compte des critères statistiques mais également de critères syntaxiques indirects (cf IV-2-1). A cet effet, le résultat d'une indexation manuelle du chapitre précédent a été confronté au contenu des cliques obtenues pour ce chapitre, et 80% des groupes nominaux résultats de l'indexation manuelle ont été effectivement retrouvés dans les cliques.

A titre d'exemple, considérons les résultats suivants où les groupements obtenus lors de l'indexation manuelle du chapitre sont écrits entre guillemets, et les cliques obtenues pour les mêmes termes sont représentées entre parenthèses.

"document technique"
(document technique service)

"document standard"
(document standard)

"document spécifique"
(document spécifique)

"document soumis à validation"
(document à validation)

"langage de programmation"
(de langage programmation)
(et langage programmation)

"structure de la documentation"
(de structure documentation)
(et structure documentation)

b- Au niveau de la recherche de documents, le contenu des cliques permet notamment des inférences lors de la reformulation des requêtes. En effet pour un terme contenu dans la base des cliques, on peut avoir accès à son environnement sémantique dans le corpus.
Par exemple pour le terme 'symbole' du chapitre considéré, on obtient l'environnement suivant:
(symbole convention)
(symbole littéral)
(symbole sigle)
(symbole graphique)

3. RESULTATS DE L'INDEXATION

A titre d'exemple, nous donnons les résultats concernant un chapitre significatif du corpus traité (il s'agit d'un chapitre suivi d'une annexe); les résultats de ce chapitre seront détaillés et suivis d'une comparaison de la méthode avec une indexation effectuée manuellement.

3.1. Evaluation quantitative de la méthode

Le chapitre analysé représente environ 3000 mots significatifs de texte et porte sur la "documentation technique" (certaines de ses caractéristiques ont été vues en VIII-2).
Pour ces 3000 mots, l'analyse syntaxique a produit 871 GCP, ce qui représente environ deux GCP pour 10 mots initiaux non vides

du texte. La taille des GCP n'excède pas 5, en raison de la limitation de la syntaxe donnée en VII-3-2 et des lacunes du traitement morphologique effectué qui empêchent la formation de GCP plus larges. Leur répartition selon les différents modèles syntaxiques est la suivante:

- 20 GCP de taille 5 correspondants aux modèles: SAAPS, SPSAA, SAPSA, SSAPS, SSPSA.
  Le pourcentage de cette catégorie par rapport au nombre total de GCP n'est que de 2,2%.

  EXEMPLE:
  (symbole graphique génér pour radiocommunication)
  (aide d annotation éventu manuscrit)
  (besoin de maintenance prévent manu)
  (programme écrit en langage clair)

- 106 GCP de taille 4 correspondants aux modèles: SAPS, SPSA, SSPS, SPSS, SSAA.
  Ce qui donne un pourcentage de 12,1% par rapport au nombre total de GCP et on remarque qu'on ne peut avoir plus de deux adjectifs associés à un substantif.

  EXEMPLE:
  (normalisation relat à documentation)
  (information contenu dans notice)
  (établissement de documentation technique)
  (opération de contrôle technique)

- 259 GCP de taille 3 corespondants en majorité au modèle SPS, avec un pourcentage de 29,7%.

  EXEMPLE:
  (contrôle en usine)
  (plan de percement)

(fonction logique réalisé)
(document de recolement)
(schéma sujet à modification)
(dessin de câblage)
(notation symbole littér)
(documentation sur microfilm)

- 114 GCP de taille 2 correspondants aux modèles: SS, SA (en majorité).
  Le pourcentage de cette catégorie (13%) est inférieur à celui de la catégorie précédente et donc la proportion de GCP d'un certain modèle n'est pas toujours inversement proportionnelle à la taille du modèle.

  EXEMPLE:
  (document NDC)
  (pièce détaché)
  (document technique)
  (document spécifique)

- 372 GCP de taille une correspondants donc au seul modèle S et avec un pourcentage de 42,7% du nombre total de GCP. On remarque que cette catégorie de GCP est largement prédominante, cela étant dû en grande partie aux lacunes de la phase de lemmatisation du vocabulaire qui entravent le bon déroulement de l'analyse syntaxique (cf VII-3).

  EXEMPLE:
  (document)
  (documentation)
  (bordereau)
  (modularité)
  (constructeur)
  (édition)
  (dossier)

La production de ces 871 GCP à partir du texte initial lemmatisé, a nécessité en temps d'exécution sous FOLL-PROLOG environ 2mn soit 0,13s par GCP. Le même texte a été analysé via l'analyseur codé en LISP, en cours d'achèvement dans le groupe, et le temps d'exécution est nettement inférieur (facteur de 4).

La base de connaissances considérée pour l'indexation de ce chapitre a été limitée dans un premier temps à la base des cliques extraites de ce même chapitre et dont les caractéristiques ont été vues en VII-2 (cette limitation permet de confronter, entre autre, le résultat de l'indexation et le contenu des cliques pour ce chapitre).
La taille initiale de cette base (en considérant le lexique des termes lemmatisés qu'elle contient), avant toute indexation, est de 79 Kmots.
Pour les 871 GCP obtenus via l'analyse syntaxique, le processus d'indexation a produit 440 termes d'indexation pour le chapitre complet.
Pour le texte initial regroupant 3000 mots, le nombre de termes d'indexation retenus est donc d'environ un terme d'indexation pour dix mots lus, en remarquant qu'il s'agit des mots non vides du texte.

Les temps d'exécution du processus d'indexation du chapitre complet (avec remontée des termes vers les niveaux supérieurs) est de 2mn 40s, ce qui donne 0,3s par GCP traité et donc 0,05 par mot initial.
Ces temps restent importants (bien qu'il s'agisse d'une indexation très fine); on pense les améliorer d'une part par une intégration des diverses fonctions de l'application, d'autre part en passant à une version compilée de ces fonctions.

Après l'indexation complète du chapitre, la taille de la base, contenant donc l'ensemble des éléments de la relation d'indexation, est de 90 Kmots.

3.2. Evaluation qualitative de la méthode

Parmi les 440 termes d'indexation retenus pour le chapitre, 38 termes sont les formes principales retenues à partir des titres. La répartition des 440 termes d'indexation, selon la syntaxe des GCP donnée en VII-3-2, est la suivante:

- 2 termes de taille 5 correspondant aux modèles: SAPSA, SAAPS, avec un pourcentage de 0,1 par rapport au nombre de GCP initiaux de même modèle et un pourcentage inférieur à 0,005 par rapport au nombre total des termes d'indexation.
  Ces pourcentages sont faibles mais prévisibles étant donné que la fréquence totale d'un GCP est souvent inversement proportionnelle à sa taille; ainsi, un GCP de taille importante a une chance minime d'être retrouvé tel quel dans une clique de la base, d'où sa décomposition en éléments plus restreints via l'opération de pattern-matching.

- 16 termes de taille 4 et de modèles: SAPS, SPSA, SSPS, SPSS. Ce qui donne un pourcentage de 71% par rapport au nombre de GCP de même modèle et 3,63% par rapport au nombre total de termes d'indexation obtenus.

- 114 termes de taille 3 dont la majorité présente le modèle SPS. Ce qui donne un pourcentage de 44,1% par rapport au nombre de GCP de même modèle et 25,9% du nombre total de termes d'indexation obtenus.

- 49 termes de taille 2 dont la majorité a le modèle SA; d'où un pourcentage de 42,9% du nombre de GCP initiaux de même modèle et 11,1% du nombre de termes d'indexation obtenus.

- 259 termes de taille 1. On obtient pour cette catégorie 69% des GCP initiaux de même modèle et un pourcentage de 58,8% par rapport au nombre total des termes d'indexation. Comme

dans le cas des GCP initiaux, ce modèle est largement prédominant au niveau des termes d'indexation.

On remarque que dans cette répartition, la catégorie de modèle SPS présente le plus grand pourcentage par rapport au nombre de GCP initiaux du même modèle (mis à part le modèle S), mais que ce pourcentage est souvent faussé du fait du processus de pattern-matching. Comme exemple le GCP initial "opération de contrôle technique" a été décomposé au seul terme d'indexation "opération de contrôle".

Les 38 termes d'indexation relatifs aux titres des entités de structure se répartissent selon la même syntaxe en:

- deux termes de taille quatre (SPSA, SAPS) et qui sont:

  (établissement de documentation technique)
  (diffusion particulier de documentation)

- 14 termes de taille 3 (le seul modèle étant SPS) et qui sont:
  (documentation de contrôle)
  (structure de documentation)
  (titre d marché)
  (diffusion de documentation)
  (réception de documentation)
  (qualité de documentation)
  (recommandation et spécification)
  (documentation de référence)
  (référence du CNET)
  (approbation de documentation)
  (établissement du bordereau)
  (contrôle sur chantier)
  (annexe I princip)
  (contrôle en usine)

- 9 termes de taille deux dont un seul présente le modèle SS et les autres le modèle SA:
  (unité employé)
  (support papier)
  (disposition contractu)
  (disposition génér)
  (principe génér)
  (document standard)
  (cas échéant)
  (document spécifique)
  (documentation région)

- 12 termes de taille 1 et qui sont:
  (symbole)
  (traitance)
  (responsabilité)
  (pénalité)
  (document)
  (jour)
  (structure)
  (support)
  (présentation)
  (format)
  (bordereau)
  (documentation)

Au niveau des titres, les lacunes du traitement morphologique se ressentent également; comme exemple, le groupe nominal "mise à jour des documents" n'a pu être retenu comme GCP, lors de l'analyse syntaxique, car il y a ambiguité sur le terme "mise" qui a été pris comme adjectif qualificatif.
On remarque que certains titres assez vides ont été retenus comme termes d'indexation; c'est le cas, par exemple du titre "principe général" qui introduit une entité ayant même aspect qu'une introduction ou un résumé (aspect général), d'où la

nécessité d'enregistrer ce titre vide dans la table des TITRES-PROCEDES.

La méthode d'indexation proposée étant fondée sur la structure logique du texte traité, il est possible de comparer les poids d'indexation associés à un terme donné, même si la base ne comporte qu'une seule unité d'indexation maximale analysée (ce qui est le cas pour ce chapitre). A titre d'exemple, on présente ci-après les résultats obtenus, pour deux entités du chapitre. Ceci est suivi d'une comparaison avec une indexation manuelle établie, pour ces deux entités, par une documentaliste.

1) <u>Première entité</u>:

Le texte associé à cette entité est le suivant:

"dispositions générales:
Les dispositions générales concernant la documentation technique relative à un marché figurent dans le "cahier des clauses administratives générales applicables aux marchés industriels passés au nom de l'Etat". Les modalités d'application aux marchés de télécommunications, quel que soit l'objet de ceux-ci: études, fournitures, installations, sont précisées ci-après et sont éventuellement complétées dans les clauses du marché et les instructions de l'Administration. La nature et le nombre des documents à fournir peuvent en effet varier considérablement selon l'objet du marché, mais le but et la structure de la documentation restent les mêmes. La documentation comprend des documents standard et des documents spécifiques. On appelle document standard (d'un produit) tout document soumis à validation en même temps que ce produit. Il ne peut être remis en question sans procéder à une nouvelle validation. Il est utilisé par tous les titulaires de marchés qui concernent ce produit. On appelle document spécifique (d'un marché déterminé) tout document qui complète les documents standard pour ce

marché. Les clauses suivantes fixent, sauf prescriptions spéciales du marché ou indications complémentaires de l'Administration, les qualités de la documentation à fournir, sa structure, sa diffusion ainsi que les délais à respecter".

Les termes d'indexation obtenus pour cette entité sont au nombre de 32, répartis selon la valeur du poids de la manière suivante (chaque terme est suivi de son poids d'indexation dans cette entité):

- Valeur du poids supérieure ou égale à 0,5:
  ((disposition génér) 1.0)
  ((objet) 0.5232558)
  ((document pour marché) 0.5116279)
  ((marché de télécommunication) 0.5116279)
  ((document à validation) 0.5116279)
  ((instruction de administration) 0.5116279)
  ((marché au nom) 0.5116279)
  ((but) 0.5116279)
  ((application au marché) 0.5116279)
  ((modalité) 0.5116279)

- Valeur entre 0,1 et 0,4:
  ((validation) 0.3565891)
  ((nature et nombre) 0.2616279)
  ((qualité de documentation) 0.2616279)
  ((état) 0.2616279)
  ((document standard et document spécifique) 0.2616279)
  ((marché) 0.2061310)
  ((indication) 0.1899224)
  ((nom) 0.1782945)
  ((temps) 0.1782945)
  ((structure de documentation) 0.1782945)
  ((document standard) 0.1782945)
  ((structure) 0 2 0.1482558)

((document spécifique) 0.1366279)
((instruction) 0.1116279)

- Valeur du poids inférieur à 0,1:
  ((application) 8.3056478e-02)
  ((étude) 8.3056478e-02)
  ((délai) 7.4127907e-02)
  ((administration) 6.7183462e-02)
  ((fourniture) 6.1627907e-02)
  ((diffusion) 6.1627907e-02)
  ((installation) 4.2877907e-02)
  ((document) 3.50205e-02)
  ((documentation) 2.2044573e-02)

Les poids d'indexation des termes retenus varient entre 1 (pour le titre qui ici est vide) et 2.2044e-02. Ils sont fonction des critères statistiques choisis mais ils peuvent être très faussés par la faiblesse de l'analyse morphologique, d'une part, et par la cassure syntaxique qui décompose un GCP large et donc précis en sous-structures plus restreintes mais fréquentes, d'autre part.

Les termes dont le poids est proche de la valeur 0,5 sont de bons termes d'indexation lorsque leur taille est importante. Plus la taille d'un terme de cette catégorie est grande, plus le terme discrimine bien l'entité où il existe; c'est le cas, par exemple, des termes "marché de télécommunication", "document pour marché", "document à validation". Par contre les termes "objet" et "but" sont des termes non significatifs affectés de poids importants: ceci est en partie dû à la cassure syntaxique d'un GCP plus large qui les contenait ("objet du marché", ...).

Les termes dont le poids est compris entre 0,1 et 0,4 sont moyennement bons car assez fréquents dans le chapitre pour discriminer totalement une entité par rapport à l'ensemble.

C'est le cas des termes "document standard et document spécifique", "structure de la documentation", "qualité de la documentation".

On remarque qu'à tout terme d'indexation de longueur supérieure à 1, est associé un poids supérieur à 0,1; par contre tout terme dont le poids est inférieur à 0,1 est un mot isolé qui caractérise le chapitre globalement mais est trop fréquent dans ce chapitre pour caractériser très précisément une entité donnée. C'est le cas des termes: "document", "documentation", "diffusion", "fourniture", "étude", "application" pour lesquels les fréquences totales respectives sont 105, 50, 16, 10...

L'indexation manuelle pour ce même texte a produit 22 termes d'indexation dont 18 ont été retrouvés dans le processus automatique. Les termes non retrouvéss ont en fait tous posé un problème d'ambiguité dans le traitement morphologique; par exemple, le terme "titulaire du marché" n'a pas été retenu comme GCP lors de l'analyse, car une ambiguité réside dans le mot "titulaire" qui a été considéré comme adjectif.

Les termes retrouvés dans le processus automatique et non retenus dans l'indexation manuelle sont au nombre de 10; ils ne sont généralement pas porteurs d'information ("nom", "temps", "objet...); ces termes sont souvent le résultat d'une cassure syntaxique d'un GCP large.

2) <u>Deuxième entité analysée</u>:

Le texte de cette entité est le suivant:

"La structure de la documentation doit permettre de localiser facilement les documents ou informations ci-après: le bordereau auquel sera adjointe la liste de diffusion des documents (services destinataires, documents diffusés, types de support,

dates contractuelles et effectives de livraison); les conventions de représentation (répertoires de termes, définitions, sigles, symboles, conventions graphiques,...), avec références aux normes et spécifications en vigueur; les règles de numérotage et de classement de la documentation (qui doivent permettre de vérifier facilement si toute série de documents est complète); les règles de désignation des conducteurs et des organes; la notice technique générale exposant la structure d'ensemble, les interfaces mises en jeu, la chronologie du fonctionnement, les dispositions de programmation. Elle sera accompagnée d'un schéma d'ensemble représentant les liaisons entre les différents sous-ensembles. Elle comprendra également sous forme de documents séparés une description des contraintes de dimensionnement du système (modularité, crans d'extension, calcul des configurations, établissement des listes de matériel) et un catalogue des équipements (matériel et logiciel); une partie de cette notice devra en outre regrouper les plans ou tables des matières des autres documents; les documents de définition ou livrets d'équipements propres à chaque type de matériel, précisant les équipements des baies, cadres ou sous-ensembles constitutifs; les documents permettant de suivre l'évolution du matériel (feuilles de filière, ou chemin de fer); les documents relatifs au logiciel: dossiers d'analyse et de programmation, organigrammes, spécifications des langages de programmation employés, avec commentaires ou programmes écrits en langage clair, et description des tables, fichiers, traitements ainsi que des moyens informatiques de maintenance et de gestion; une table récapitulative des différents sous-ensembles homogènes du logiciel (modules de logiciel) avec leur identification codée; les schémas de principe de tous les sous-ensembles et équipements, et les notices de fonctionnement qui leur correspondent; les dessins de câblage et les notices s'y rapportant; les documents nécessaires à l'implantation, à l'installation et au raccordement: valeurs locales des paramètres de configuration, diagramme des liaisons, plans de

salle (avec adresse et plan de situation), plan d'emplacement de chacun des sous-ensembles ou équipements, câblage entre baies ou organes, schémas des alimentations en énergie et des protections, repérage des bornes et connexions, dessins de construction avec cotes d'encombrement, plans de percement, plans de sectionnement, ...; les dossiers définissant les procédés de vérification ou de réception de tout ou partie des fournitures ou installations; les spécifications des matières premières et des pièces; les spécifications de réglage des appareils et organes; les données relatives aux composants; la documentation d'exploitation et de maintenance comprenant : un manuel d'exploitation structuré par ensembles d'opérations de même type (commandes de gestion des abonnés, de gestion de la traduction, ...) comportant une notice par commande du langage de relations homme-machine. Chaque notice donnera la description fonctionnelle de l'opération et les interventions sur le matériel associées, la syntaxe et la sémantique de la commande, les résultats escomptables et la liste des messages d'erreurs; un manuel de maintenance décrivant :

l'organisation du logiciel de maintenance du système et les commandes du langage de relations homme-machine associées;

pour chaque type d'équipement, les besoins de maintenance préventive manuelle avec mention de périodicité recommandée pour les essais et des mesures à prendre en fonction des résultats obtenus et les actions de maintenance corrective (localisation des fautes et dépannage). Les précautions à prendre pour la sécurité du personnel et des matériels seront précisées ainsi qu'éventuellement les instructions de démontage et de remontage pour l'accès à certains organes;

les dessins de l'outillage nécessaire à la maintenance;

les possibilités de surveillance locale et de télésurveillance; un dictionnaire de fautes, donnant la liste des messages de faute et d'alarme classés selon une règle simple et indiquant, pour chaque message, sa signification ainsi que les interventions nécessaires sur le matériel et le logiciel (en

faisant référence, si nécessaire, au manuel de maintenance); les nomenclatures, listes ou catalogues des composants (complétées s'il y a lieu, par les tables de vérité des fonctions logiques réalisées par ces composants) ou pièces détachées entrant dans la composition du matériel, avec toutes références de commande nécessaires; les plans de brassage, les combinaisons des traducteurs, les étiquettes, et plus généralement les plans ou schémas sujets à des modifications en cours de chantier".

Les termes d'indexation obtenus à ce niveau sont au nombre de 159, répartis selon leur poids de la manière suivante:

- valeur supérieur à 0,5:
  ((langage de relation homme) 0.5044247)
  ((baie ou organe) 0.5022123)
  ((gestion de traduction) 0.5022123)
  ((spécification de réglage) 0.5022123)
  ((ensemble ou équipement) 0.5022123)
  ((ensemble et équipement) 0.5022123)
  ((documentation d exploitation) 0.5022123)
  ((dictionnaire de faute) 0.5022123)
  ((convention de représentation) 0.5022123)
  ((commande du langage) 0.5044247)
  ((message de faute) 0.5022123)
  ((série de document) 0.5022123)
  ((document de définition) 0.5022123)
  ((schéma d ensemble) 0.5022123)
  ((partie de notice) 0.5022123)
  ((plan de situation) 0.5022123)
  ((maintenance du système) 0.5022123)
  ((commande de gestion) 0.5022123)
  ((type d équipement) 0.5022123)
  ((document diffusé) 0.5022123)
  ((maintenance manu) 0.5044247)

((date contractu) 0.5022123)
((équipement propre) 0.5022123)
((exploitation structuré) 0.5022123)
((document nécessaire) 0.5022123)
((document relat) 0.5022123)
((fer) 0.5022123)
((convention) 0.5022123)
((implantation) 0.5022123)
((classement) 0.5022123)
((emplacement) 0.5022123)
((construction) 0.5044247)
((fichier) 0.5022123)
((paramètre) 0.5022123)
((filière) 0.5022123)
((livraison) 0.5022123)
((évolution) 0.5022123)
((raccordement) 0.5022123)
((analyse) 0.5022123)
((énergie) 0.5044247)
((numérotage) 0.5022123)
((module) 0.5022123)
((chemin) 0.5022123)
((configuration) 0.5044247)
((identification) 0.5022123)
((borne) 0.5022123)
((remontage) 0.5022123)
((périodicité) 0.5044247)
((démontage) 0.5022123)
((traducteur) 0.5022123)
((connexion) 0.5022123)
((câblage) 0.5066371)
((mesure) 0.5022123)
((organigramme) 0.5022123)
((besoin) 0.5022123)
((essai) 0.5044247)

((adresse) 0.5022123)
((vérité) 0.5022123)
((signification) 0.5022123)
((intervention) 0.5044247)
((télésurveillance) 0.5022123)
((alarme) 0.5022123)
((traitement) 0.5022123)
((catalogue) 0.5044247)
((repérage) 0.5022123)
((combinaison) 0.5022123)
((erreur) 0.5022123)
((machine) 0.5044247)
((alimentation) 0.5022123)
((pièce) 0.5044247)
((abonné) 0.5022123)
((description) 0.5066371)

Comme c'était le cas sur l'exemple précédent, les termes d'indexation ayant un poids supérieur à 0,5 sont de bons termes d'indexation lorsque leur taille est importante: dans cette catégorie seuls 6 termes sur 26 sont d'un caractère général ("série de documents", "type d'équipement", "partie de notice", ...).
Parmi les termes isolés ayant un poids important, certains sont caractéristiques de l'entité traitée ("télésurveillance", "alarme", "remontage, "démontage", ...) et d'autres sont des mots outils ("besoin", "essai", ...).

- Valeur supérieure à 0,3:
  ((table) 0.4088495)
  ((dessin) 0.3816371)
  ((commande) 0.3421828)
  ((matière) 0.3377581)
  ((fonction) 0.3377581)
  ((mention) 0.3377581)

((message) 0.3377581)
((résultat) 0.3377581)
((règle) 0.3066371)

- Valeur supérieure à 0,2:
((cote d encombrement) 0.2522123)
((livret d équipement) 0.2522123)
((schéma de principe) 0.2522123)
((fourniture ou installation) 0.2522123)
((langage de programmation) 0.2522123)
((norme et spécification) 0.2522123)
((notice de fonctionnement) 0.2522123)
((plan de percement) 0.2522123)
((plan de sectionnement) 0.2522123)
((répertoire de terme) 0.2522123)
((a lieu) 0.2522123)
((baie) 0.2044247)
((possibilité) 0.2522123)
((type) 0.2044247)
((protection) 0.2522123)
((ensemble) 0.2132743)
((programme) 0.2544247)
((vérification) 0.2522123)
((valeur) 0.2522123)
((commentaire) 0.2522123)
((spécification) 0.2544247)
((appareil) 0.2522123)
((surveillance) 0.2522123)
((cadre) 0.2522123)
((maintenance) 0.286001)
((désignation) 0.2522123)
((organisation) 0.2522123)
((cote) 0.2522123)
((programmation) 0.2544247)
((sigle) 0.2522123)

((feuille) 0.2522123)
((liste) 0 4 0.2588495)
((dossier) 0.2544247)
((opération) 0.2945638)

- Valeur supérieure à 1:
((type de support) 0.1272123)
((structure de documentation) 0.1688790)
((plan de salle) 0.1272123)
((liste de diffusion) 0.1688790)
((spécification en vigueur) 0.1688790)
((plan de brassage) 0.1688790)
((diagramme) 0.1022123)
(notice) 0.1421828)
((gestion) 0.1272123)
((extension) 0.1688790)
((instruction) 0.1022123)
((norme) 0.1688790)
((système) 0 2 0.1294247)
((définition) 0.1710914)
((fonctionnement) 0.1022123)
((référence) 0.143000)
((plan) 0.1040876)
((modification) 0.1688790)
((livret) 0.1638790)
((faute) 0.1688790)
((organe) 0.143000)
((étiquette) 0.1688790)
((calcul) 0.1688790)
((liaison) 0.1710914)

Parmi ces termes, certains sont représentatifs de l'entité mais également d'autres entités et n'ont donc pas un bon pouvoir discriminant (poids moyen); c'est surtout le cas des termes de taille importante ("livret d'équipement",

"structuration de la documentation", "langage de programmation", "plan de brassage"...).

- Valeur inférieure à 0,1:
  ((documentation) 1.2629056e-02)
  ((chantier) 6.4712389e-02)
  ((partie) 3.5545723e-02)
  ((équipement) 0.07585335)
  ((installation) 3.3462389e-02)
  ((langage) 6.4712389e-02)
  ((service) 5.2212389e-02)
  ((disposition) 8.5545722e-02)
  ((schéma) 2.8528179e-02)
  ((structure) 6.4712389e-02)
  ((document) 4.0473711e-02)
  ((symbole) 8.5545722e-02)
  ((bordereau) 3.3462389e-02)
  ((établissement) 0.07364096)
  ((information) 5.7767945e-02)
  ((réception) 8.5545722e-02)

Cette catégorie regroupe les termes isolés qui sont génériques pour le chapitre, et dont le poids est par conséquent très faible ("document", "documentation", "information", "bordereau"...).

L'indexation manuelle de cette entité a fourni 145 termes d'indexation soit 14 termes de moins par rapport à l'indexation automatique. Ceci est dû, d'une part aux lacunes du traitement morphologique, et d'autre part à l'inexistance de groupes dans les cliques. A ce niveau, on remarque qu'un grand pourcentage des termes de cette catégorie (42,8%) sont décomposés et que la totalité de leurs composants sont retenus en tant que termes d'indexation avec un poids identique (comme exemple, "langage de relation homme-machine" a été décomposé en "langage de relation

homme" et "machine").
Dans cette catégorie, on remarque également que les adjectifs sont souvent supprimés dans les groupes d'indexation (30,3%). (comme exemple "fonction logique" donne "fonction", "maintenance corrective" donne "maintenance"...).
Le nombre de termes choisis manuellement et non retrouvés complètement est de 15.

Le nombre de termes retenus dans l'indexation automatique et non dans l'indexation manuelle est de 31. Dans cette catégorie, certains sont vides ("besoin", "essai", "partie" ...) mais d'autres sont généraux pour le chapitre ("livret", "module", "commande", "norme et spécification" ...).

Certains termes retrouvés automatiquement ne figurent pas parmi les termes d'indexation manuelle bien qu'ayant un poids assez correct (EX: "exploitation structurée", "commande de gestion", "type de support"...).

3.3. Indexation dynamique

Nous avons expérimenté ici l'efficacité de la stratégie de remontée de termes d'indexation depuis les entités minimales vers les entités maximales.
Etant donnée la stratégie proposée qui consiste à ne considérer au niveau de l'entité père que l'ensemble des termes communs aux entités filles, le processus est très sensible à la finesse de structuration du document. Plus le degré moyen de l'arborescence est élevé, moins il y a de chances pour qu'il existe des termes communs à un niveau donné, et donc moins il y a de chances pour que des termes remontent au niveau supérieur. Ceci est particulièrement sensible pour un document comme les NEF, dont le niveau de structuration est très poussé (certaines entités sont réduites à une phrase de spécification). Afin de souligner cette incidence, nous avons donc considéré successivement deux

niveaux d'indexation minimaux (le sous paragraphe et le sous chapitre), le niveau d'indexation maximal restant le document, constitué du chapitre 15 des NEF et de son annexe. Pour illustrer le résultat de la méthode, nous avons considéré le comportement d'un terme particulier "document", à travers les différents niveaux d'indexation.

a) Niveau d'indexation minimal égal au sous-chapitre

Deux termes remontent au niveau racine de la hiérarchie du document: "document" et "documentation", avec des poids respectifs de 0,59 et 0,04. La figure ci-dessous montre l'évolution de la mesure asociée au terme "document" le long de la hiérarchie de ce chapitre:

Niveau	Poids
(document)	0,59
(chapitre)	0,27
(sous chapitre1)	0,11
(sous chapitre2)	0,063
(sous chapitre3)	0,078
(sous chapitre4)	0,057
(annexe)	0,50
(sous chapitre)	1

Cet exemple illustre bien le comportement de "document" en tant que terme générique: le poids d'indexation du terme est strictement décroissant dans le sous arbre (document (chapitre (sous chapitre))) donné dans l'ordre préfixé de parcours. Le poids maximal est obtenu pour le niveau sous chapitre de l'annexe, où le terme figure dans le titre. Dans un processus d'interrogation, on donnerait donc en priorité l'annexe du chapitre 15, et en second choix l'ensemble du chapitre 15.

Il est également intéressant de considérer le comportement

des termes d'indexation comportant "document" dans ce processus. Si on examine, par exemple, l'entité (sous chapitrel), on trouve la répartition suivante des termes d'indexation et de leur poids:

Termes	Poids
(document sujet)	0,17
(document contractuel)	0,54
(document standard)	0,17
(document spécifique)	0,127
(document pour marché)	0,50
(document nécessaire à exécution)	0,50
(document soumis à approbation)	0,50
(document standard et document spécifique)	0,50

Tous ces termes ont un poids supérieur à celui de "document" dans l'entité considérée. Ils sont davantage précis et caractéristiques du contenu du sous chapitre, que le terme isolé.

b) <u>Niveau d'indexation minimal égal au sous-paragraphe</u>

Etant donnée la structuration extrêmement fine du texte, le niveau de remontée des termes est très limité, pour les raisons évoquées ci dessus. Pratiquement, la remontée est limitée au niveau paragraphe, et pour cinq d'entre eux seulement. Si cela est normal dans le contexte particulier de l'expérimentation, on peut également en déduire un raffinement dans la stratégie de remontée des termes:
dans le cas où aucun terme commun n'est trouvé dans les entités filles, casser les groupes d'indexation pour obtenir des termes plus généraux, et relancer le processus sur l'ensemble de ces sous termes.

3.4. Conclusion

Cette expérimentation sur une portion réduite du corpus fait apparaître la qualité de la méthode (précision identique à celle d'une indexation manuelle fine), mais des lacunes subsistent au niveau des traitements linguistiques initiaux (analyse morphologique). Ces défauts devraient disparaître avec la version ultérieure de l'analyseur. Un autre défaut apparent est le nombre élevé de termes produits, et donc l'encombrement de cet ensemble d'informations. Cet aspect est évidemment lié à la finesse d'indexation recherchée qui fournit la diversité des termes.

Pour un corpus figé, limité aux NEF, on peut évidemment envisager une sélection a posteriori dans cet ensemble de termes. Pour un corpus évolutif, il faut renforcer la sélectivité de la base de connaissances en restreignant la définition des cliques selon des critères plus sévères dans la mesure d'associations entre termes: on obtient alors moins de cliques en proportion, et leur cardinal moyen est plus faible, ce qui entraine un taux de cassure plus élevé des GCP, et donc une meilleure factorisation.

CHAPITRE IX

CONCLUSION

L'indexation est l'opération clé dans un système de recherche d'informations: elle sert de pont entre le langage des documents et celui des questions. C'est sur sa qualité que repose l'efficacité du système.

L'information traitée étant essentiellement un contenu exprimé en langue naturelle, l'opération d'indexation hérite des complexités et des ambiguités d'interprétation de la langue naturelle et relève de l'étude des liaisons syntaxe-sémantique. Du point de vue informatique, ce domaine a suscité beaucoup d'intérêt mais reste jusqu'à présent un problème très ardu à traiter de manière automatique: relevant, entre autre de la pragmatique et du subjectif, il est difficilement maitrisable par machine.

Parmi les systèmes d'indexation existants, certains exploitent à divers niveaux, des éléments purement sémantiques qui généralement sont construits manuellement et donc à des coûts élevés; ils posent de plus le problème classique de cohérence et de complétude par rapport au domaine effectivement couvert par le corpus.

Dans la méthode d'indexation proposée, nous avons tenté d'apporter notre contribution à la résolution de ce problème en nous intéressant tout d'abord aux aspects structurels du document qui renseignent sur sa sémantique, et nécessitent peu d'a priori pour leur exploitation:

- l'aspect structurel global du document, a permis de définir une stratégie générale d'indexation tenant compte d'une interprétation différenciée des divers types de composants

textuels (titres, textes spéciaux...), et aboutissant à une procédure simple d'indexation dynamique à travers toute la hiérarchie du document.

- l'aspect syntaxique de la langue naturelle qui a permis de redéfinir la notion d'élément d'indexation en termes de syntagmes, tout en contrôlant l'explosion combinatoire associée par une normalisation via une base de connaissances établie automatiquement.

Nous avons enfin montré comment on pouvait parvenir à un plus juste équilibre entre la part enregistrée et la part calculée de la relation d'indexation, en limitant l'enregistrement aux termes d'indexation explicitement trouvés dans le texte, et en laissant l'évaluation (lors de l'interrogation) des termes sous-jacents non effectivement rencontrés.

Sur le plan des résultats pratiques, il est clair qu'on ne pourra vraiment juger de l'efficacité de la méthode qu'au travers de l'exploitation de ses résultats via un système d'interrogation. Les études comparatives pour quelques chapitres importants du corpus traité, entre l'indexation manuelle et l'indexation automatique montrent cependant des résultats très encourageants sur le plan qualitatif. Sur le plan quantitatif, les performances enregistrées sont encore insuffisantes, bien qu'à ce niveau, il faudrait comparer le coût d'une indexation manuelle aussi fine pour obtenir une appréciation plus exacte.

Les logiciels présentés sont encore largement expérimentaux, pas intégrés, ce qui explique en grande partie ces performances; notre premier souci ayant été d'évaluer tout d'abord les performances qualitatives de la méthode.
Des progrès significatifs peuvent être réalisés de ce coté par une meilleure intégration des outils dans un langage unique plus

performant ce qui fournira les conditions d'exécution et d'enchainement des opérations. Cette étude est actuellement en cours dans le groupe, à travers le langage LELISP [CHA 84] sur VAX 780 (sous le système UNIX), dont on attend beaucoup quant aux possibilités de compilation et de mise en oeuvre de processus parallèles.

Sur le plan qualitatif, il est clair que les améliorations les plus importantes relèvent du domaine linguistique. Notre analyse a été souvent perturbée par l'absence d'une analyse morpho-syntaxique efficace, capable notamment de résoudre le problème d'homographie.

Nous comptons donc, dans un avenir proche intégrer l'analyseur morpho-syntaxique [PALM 81] qui est en cours de réalisation dans le groupe. Cet analyseur permettrait d'expérimenter la méthode avec des éléments d'indexation plus précis et plus complets (cf IV-1). Du point de vue des performances, cet analyseur présente deux avantages: de part sa conception basée sur une méthode d'apprentissage au niveau de la morphologie et au niveau des règles syntaxiques, il permet un traitement entièrement automatique (après une légère initialisation du dictionnaire).

De manière générale, nous pensons que le problème des performances quantitatives de ce type de logiciel se posera toujours, étant donné la masse des informations traitées et la complexité des traitements linguistiques et statistiques. Comme dans d'autres domaines, il nous paraît clair que de substantielles améliorations pourront être apportées par le transfert de certaines fonctions logiciel de base vers le matériel (microprogrammation). C'est le cas notamment de certaines fonctions linguistiques comme l'analyse morphologique et la gestion du dictionnaire; ce transfert est à l'étude dans notre groupe.

En ce qui concerne une prise en compte des différents types d'objets existants dans un document, l'intégration concernerait un outil syntaxique généralisé qui permettrait la manipulation d'objets "multi_syntaxe". Un logiciel tel que MENTOR [LAN 81], [LAN 83], [MEL 83], [LAN 81], [LAN 83], [MEL 83], possède ces potentialités, et permettrait de prendre en compte non seulement la structure hiérarchique d'un document, mais aussi celles d'objets non textuels (comme les programmes ou les figures) pour lesquels une définition syntaxique est possible. Bien que nous étant intéressé pour l'instant aux informations purement textuelles, cette possibilité nous parait intéressante pour intégrer ultérieurement différentes stratégies d'indexation propres à des données de nature différente.

BIBLIOGRAPHIE

(ADA 81)
ADAM A.; 'MENTOR: bilan et perspectives des éditeurs de programmes'; dans "structures de l'informations"; num 24 sur COLLOQUE 'Intelligence artificielle'; TOULOUSE juillet 81.

(AND 73A)
ANDREEWSKY A., FLUHR C.; RAMBOUSEK J.;
'Automatisation de l'analyse discriminante, de l'indexation, de la recherche hiérarchisée des documents et de l'aide à la décision';
NOTE CEA-N-1650 SACLAY juin.1973

(AND 73B)
ANDREEWSKY A.; FLUHR C.; 'Indexation automatique, maintenance et gestion d'un système documentaire, 1ere partie: aspects théoriques.';
NOTE CEA-N-1694 SACLAY déc.1973

(AND 75)
ANDREEWSKY A.; COMBRISSON F., FLUHR C.; 'Le problème de l'identification automatique des concepts';
NOTE CEA-N-1816 sep.1975

(AND 82A)
ANDRE E.; 'Journées CONCERTO: présentation des objectifs du projet';
Journées CONCERTO nov.1982

(AND 82B)
ANDRE E.; 'présentation du projet pilote CONCERTO aux journées BIGRE';
Journées BIGRE j.1982

(BCH 81)
BRUANDET M.F., CHIARAMELLA Y.; 'La recherche documentaire dans une base de données textuelles';
Bulletin du centre de Hautes Etudes Internationales d'Informatique Documentaire (CID). L'informatique documentaire 4 trimestre 1981

(BCK 82)
BRUANDET M.F., CHIARAMELLA Y., KERKOUBA D.; 'Méthodes d'indexation automatique de documentations techniques dans le cadre d'un atelier de logiciel';
JOURNEES D'ETUDES CONCERTO PERROS-GUIREC, 16-17 décembre 1982

(BCK 83)
BRUANDET M.F., CHIARAMELLA Y., KERKOUBA D.; 'Méthodes empiriques de construction de thésaurus : Expérimentation';
Bulletin du CID 1983 Premier trimestre

(BOO 75)
BOOKSTEIN A., SWANSON D.R.; 'A décision theoric foundation for indexing';
Journal of the American Society for Information Science jan-feb 1975

(BRU 80A)
BRUANDET M.F.; 'Framework for automatic and dynamic thesaurus updating in information retrieval systems'; Proceeding of Coling 80 TOKYO.

(BRU 80B)
BRUANDET M.F.; 'A propos de la construction automatique d'un thésaurus dans un système de recherche d'information'; Rapport de recherche IMAG- num 229 1980.

(BRU 81)
BRUANDET M.F.; 'Notion de concept pour la construction automatique d'un Thésaurus évolutif';
Congrès AFCET-INFORMATIQUE. Actes du congrès de l'AFCET GIF SUR YVETTE 18-20 novembre 1981

(BRU 82)
BRUANDET M.F.; 'Concept notion for automatic and dynamic thesaurus updating';
International conference on systems documentation; jan 82 SIGOA-SIGDOC LOS ANGELES.

(CHA 84)
CHAILLOUX J.; 'LELISP de l'INRIA manuel de référence version 14' Rapport I.N.R.I.A 1984

(COO 78)
COOPER W.S., MARON M.E; 'Foundations of probabilistic and utility theoric indexing'; Journal of the ACM, Vol. 25,No. 1, january 1978,p.67-80 .

(COU 77)
COURTIN J.; 'Algorithmes pour le traitement interactif des langues naturelles';
Thèse d'état USMG; GRENOBLE 1977.

(COY 72)
COYAUD M.; 'Linguistique et documentation';
"Langue et langage" LAROUSSE 1972.

(DAV 77)
DAVIS R., KING J.; 'An overview of production systems'; Machine intelligence Vol 8 p.300-332 1977 .

(DEB 77)
DEBILI F.; 'Traitements syntaxiques utilisant des matrices de précédence fréquentielles construites automatiquement par apprentissage';
Thèse de docteur ingénieur; Université de Paris VII sept.1977.

(DEB 82)
DEBILI F.; 'Analyse syntaxico-sémantique fondée sur une acquisition automatique de relations lexicales-sémantiques';
Thèse d'état; Université Paris XI Centre d'ORSAY janv.1982

(DEF 84)
DEFUDE B.; 'Knowledge based systems versus thesaurus: an architecture problem about expert systems design';
In Research and development in information retrieval;
Cambridge University Press 1984.

(DES 82)
DESCLES J.P.; 'Langages quasi-naturels articulés avec une base de connaissances: présentation et problèmes'.
Colloque "TRAITEMENT AUTOMATIQUE DES LANGUES NATURELLES ET SYSTEMES DOCUMENTAIRES" Clermond-Ferrand mai 1982.

(DEW 81)
DEWEZE A.; 'Réseaux sémantiques';
Thèse d'état Université CLAUDE BERNARD LYON 1981.

(DON 83)
DONZ Ph.; 'FOLL une extension au langage PROLOG'
manuel d'utilisation, CRISS, U2 Grenoble 1983

(FLU 77)
FLUHR C.; 'Algorithmes à apprentissage et traitement automatique des langues';
Thèse d'état; Université Paris-sud centre d'ORSAY 1977

(FON 82)
FONDIN H.; 'Le titre comme élément de description du contenu d'un

document; Recherche sur les méthodes d'évaluation';
Documentaliste, Vol 19, num 1, jan-fev 1982

(GOL 72)
GOLLEM: manuel d'utilisation SIEMENS 1972.

(GUE 82)
LE GUERN M.; 'Les descripteurs dans un système documentaire';
Colloque "traitement automatique des langues naturelles et systèmes documentaires"; Clermont-Ferrant mai 1982.

(GRO 75)
GROSS M.; 'Méthodes en syntaxe'; HERMANN PARIS 1975.

(HAM 81)
HAMEON J.; 'Indexation et classement en bureautique';
Thèse de 3ième cycle INPG GRENOBLE jan.1981.

(JOL 81)
JOLOBOFF V., LOPEZ M., KOWARSKI I.; 'Proposition pour la définition d'un système intégré de gestion de documents'; IMAG rapport interne jul 1981 .

(KAH 81)
KAHN G.; 'FLIP: manuel d'utilisation';
Rapport de recherche num 2 INRIA j.81

(KBC 82)
KERKOUBA D., BRUANDET M.F., CHIARAMELLA Y.; 'Analyse des NEF (Tome 1)'; RAPPORT de fin de contrat. PROJET CONCERTO. novembre 1983

(KER 81)
KERKOUBA D.; 'Incidence du thésaurus dans les systèmes documentaires';
Rapport de D.E.A INPG, GRENOBLE 1981.

(LAN 81)
LANG B., DONZEAU-GOUGE V., KAHN G., MELESE B.;'Intégration de MENTOR dans un poste de travail'; journées CONCERTO dec 1981.

(LAN 83)
LANG B.;'Some ideas about gates'; I.N.R.I.A rapport interne jan 83 .

(MAC 82)
'Manuel d'utilisation MACLISP sous MULTICS'
Documentation CICG 1982

(MAN 83)
MANIEZ J.; 'Problèmes de syntaxe dans les systèmes de recherche documentaire'; Documentaliste; Vol 20, num 2 mars-avril 1983.

(MAR 77)
MARON M.E.; 'On indexing, retrieval and the meaning of about'; Journal of ASIS jan 1977

(MEL 80)
MELESE B.; 'Manipulation de programmes pascal au niveau des concepts du langage';
Thèse de 3ieme cycle PARIS 11 1980

(MEL 83)
MELESE B.;'Manipulation de textes structurés sous MENTOR'; I.N.R.I.A rapport interne janvier 1983 .

(MER 82)
MERLE A.; 'Un analyseur pré-syntaxique pour la levée des ambiguités dans des documents écrits en langue naturelle: application à l'indexation automatique';
Thèse 3ieme cycle INPG 1982.

(MIS 78)
MISTRAL: manuel d'utilisation; CII-HB; La Documentation française 1978.

(PAL 81)
PALMER P.; 'Etude de l'organisation d'un dictionnaire pour l'analyse du français'; RAPPORT de DEA INPG GRENOBLE 1981.

(PAS 72)
PASSAT: manuel de description SIEMENS 1972.

(QUI 81)
QUINT V.; 'TITUS: un système pour l'édition interactive de formules mathématiques'
Rapport interne IMAG juin 1982

(RIJ 79)
VAN RIJSBERGEN C.J.; 'Information retrieval';

Second édition, BUTTERSWORTH LONDON ENGLAND 1979.

(SAL 71)
SALTON G.; 'The SMART Project.';
Prentice Hall 1971.

(SAL 75)
SALTON G.; 'Dynamic information and library processing';
Prentice Hall 1975.

(SAL 81)
SALTON G.; 'a blueprint for automatic indexing '; ACM SIGIR forum, vol. 16, No. 2, FALL 1981, p.22-38 .

(SAL 83)
SALTON G., McGILL M.J., 'introduction to modern information retrieval'; mcgraw hill book company, NEW YORK, 1983 .

(SAT 74)
BOURELLY L., CHOURAQUI E.; 'Le système documentaire SATIN1: description et manuel d'utilisation';
CNRS 1974

(SPA 71)
SPARCK-JONES K.; 'Automatic keyword classification for information retrieval';
Butterworths LONDON 1971.

(STA 79)
IBM; 'STAIRS/VS: reference manuel';
IBM system manuel 1979.

(SYNT 70)
BELLY N., BORILLO A., VIRBEL J., SIOT-DECAUVILLE N.;
'Procédures d'analyse sémantique appliquées à la documentation scientifique';
Gauthier-Villars 1970

(TIG 82)
BOGO G., RICHY H., VATTON I; 'Proposition de modèle pour la normalisation des documents';
Projet TIGRE IMAG; Rapport interne oct.1982.

(TIG 83)
BOGO G., RICHY H., VATTON I.; 'Proposition de modèle pour la

normalisation des documents';
Projet TIGRE IMAG; Rapport interne oct.1983.

(VIR 81)
VIRBEL J.;'La composante matérielle des structures textuelles';
école d'été 1981 TOULOUSE .

(VIR 82A)
VIRBEL J.; 'Contribution des recherches linguistiques à la compréhension du langage naturel'
Intellectica num 2 1982 p 15-20.

(VIR 82B)
VIRBEL J.; 'Une composante linguistique dans la représentation du document intelligent';
Communication à la journée d'étude du projet KAYAK "Bureautique et Intelligence Artificielle"; INRIA fév 1982.

(WAL 79)
WALL R.A.; 'Intelligent indexing and retrieval: a man-machine partnership';
Information Processing and Management Vol 16, pp 73-90.

(WUS 81)
WU H., SALTON G.; 'A comparison of search term weighting: Term relevance vs. inverse document frequency'; ACM SIGIR forum, Vol. 16, No. 1, Summer 1981,p. 30-39 .

ANNEXE

ALGORITHMES DES PRINCIPALES FONCTIONS DU MODULE D'INDEXATION

Avant de décrire ces fonctions, il est nécessaire de décrire les types de données qui sont manipulées:

1) TCHAR: chaine de caractères.

2) ELMEX: élément lexique.
   - élément: chaine; /*mot*/
   - cat : caractère; /*catégorie lexicale*/
   Un élément de ce type est un mot du lexique suivi de sa catégorie grammaticale.

   (Type groupe conceptuel primaire)
3) TGCP: ensemble d'éléments de type ELMLEX.
   Un élément de type TGCP à la structure d'un GCP.

   (Ensemble de groupes conceptuels primaires).
4) EGCP: ensemble d'éléments de type TGCP.

   (Type groupement d'indexation).
5) TGI: mot structuré contenant
   - numcl: entier; /*numéro de clique*/
   - gcf : TGCP; /*groupe d'indexation*/
   Un élément de ce type est un GCP de la base précédé du numéro de clique qui le contient.

6) EGI: ensemble d'éléments de type TGI.

7) TREL: structure de trois éléments.
   - freloc: entier; /*fréquence locale d'un terme d'indexation*/
   - ref : entier; /*référence à une unité textuelle*/
   - poids : réel; /*valuation d'un élément de la relation

d'indexation*/

Un élément de ce type décrit partiellement un élément de la relation d'indexation pour un terme donné et une entité textuelle donnée: poids associé à l'élément, entité référencée et fréquence locale du terme d'indexation dans cette entité.

8) EREL: ensemble d'éléments de type TREL.
Un élément de ce type décrit le sous-ensemble de la relation d'indexation correspondant à un même GCF.

9) TSSCL: structuré en trois éléments:

```
-gcf      : TGCP;/*terme d'indexation*/
-frectot : entier; /*fréquence totale du terme
                     dans le corpus*/
-relindex: EREL; /*description des éléments de
                  la relation contenant 'gçf'*/
```

C'est la définition complète du sous ensemble de la relation d'indexation correspondant à GCF.

10) ESSCL: ensemble d'éléments de type TSSCL;
C'est l'ensemble de la relation d'indexation.

11) TADJECT:

```
-groupe: ensemble d'éléments de type ELMLEX;
-numcl: entier;
```

Un élément de ce type regroupe un ensemble d'adjectifs précédés du numéro de clique qui les contient. L'introduction de ce type particulier est nécessaire à l'exposé de l'algorithme de la fonction pattern-matching, qui comprend un traitement particulier des adjectifs.

12) EENT: ensemble des entiers.

Les algorithmes ci-dessous sont présentés dans l'ordre préfixé de l'arbre d'appel correspondant à la structure modulaire choisie.

```
FONCTION PATTERN (gcp : TGCP) ----> EGI
    /* pour un GCP lu, PATTERN renvoie l'ensemble d'éléments de type
    TGI qui en découlent lorsque le pattern-matching a réussi, sinon
    renvoie vide */

DEBUT
    result := PAT(GCP)
      /* PAT renvoie un GCF si le GCP appartient au thésaurus, sinon
      vide */

    /* si le pattern réussit avec GCP, result contient le GCF qui en
    découle et le traitement est fini, sinon result contient vide */
    si result = vide
    alors
      si 'a E MODELE(GCP)
          /* 'a représente la catégorie adjectif et MODELE(GCP) est
          l'ensemble des catégories des mots du GCP */
      alors
        result := RESDYNAM(GCP)
          /* RESDYNAM(GCP) renvoie l'élément de type TGI extrait de GCP
          via le réseau dynamique, sinon vide */
      finsi
      /* dans le cas où GCP contient des adjectifs, result contient le
      GCF qu'on peut extraire avec le plus grand sous-ensemble
      d'adjectifs, sinon result contient vide */
      si result = vide
      alors
        result := RESTAT(GCP)
          /* RESTAT(GCP) renvoie l'ensemble des éléments de type TGI
          extraits du GCP initial via le réseau syntaxique, sinon
          renvoie vide */
      finsi
    finsi
    /* result contient l'ensemble des GCF associé au GCP si la
    décomposition a réussi, sinon result contient vide */
    ----> result
FIN
```

```
FONCTION RESDYNAM (gcp : TGCP) ----> TGI
    /* pour un gcp contenant des adjectifs, RESDYNAM renvoie l'élément
    de type TGI qu'on peut extraire uniquement avec un jeu d'adjectifs,
    sinon RESDYNAM renvoie vide */
DEBUT
    gcpos := OSSATURE(gcp)
      /* OSSATURE renvoie le gcp initial épuré de ses adjectifs */
    si CONTCLIQ(gcpos) = vide
      /* CONTCLIQ renvoie l'ensemble des numéros de cliques qui
      contiennent gcpos */
    alors
      result := vide
        /* si le pattern ne réussit pas avec gcpos, on ne pourra jamais
        extraire un gcf du gcp initial uniquement avec le jeu
        d'adjectifs ; result est à vide et le traitement s'arrête */
    sinon
      adject := PATADJECT(gcp,CONTCLIQ(gcpos))
        /* PATADJECT renvoie l'ensemble contenant les adjectifs qui ont
        "matchés" précédé du numéro de clique à mettre à jour */
      result := REFORMG(gcp,adject)
        /* REFORMG renvoie le gcf extrait de gcp en considérant adject
        */
    finsi
    /* si gcpos matche, result contient le gcf qu'on peut obtenir avec
    le sous-ensemble maximum d'adjectifs pour lequel le pattern réussit
    ; ce sous-ensemble pouvant être vide, le gcf peut se réduire à
    gcpos */
    ----> result
FIN
```

```
FONCTION PATADJECT (gcp : TGCP, ecl : EENT) ----> TGI
    /* gcp étant le gcp initial,  ecl l'ensemble des numéros de cliques
    qui contiennent son squelette (gcp épuré de ses adjectifs),
    PATADJECT renvoie le plus grand gcf  extrait de gcp et le numéro de
    la première clique qui le contient. Ce gcf est formé en considérant
    le plus grand sous-ensemble d'adjectifs contenus dans le gcp et
    pour lequel le "pattern-matching" a réussi */
DEBUT
    listeadj := RECLIPAD(gcp)
       /* RECLIPAD renvoie l'ensemble des adjectifs du gcp */
    adject.groupe := vide
    adject.numcl := 0
      /* adject  de type TADJECT  contient le plus grand sous-ensemble
      d'adjectifs qui "matche" */
    trouvé := faux
      /*  le  booléen  trouvé  est vrai  lorsque  le  sous-ensemble  de
      cardinalité maximale a été formé */
    elmtl := PREM(listeadject)  /* premier élément de listeadject */
    tantque (listeadject ≠ vide) et (non trouvé)
    faire
      /* lorsque trouvé est vrai, le plus grand sous-ensemble est formé
      dans adject */
      icl := LIREELMTl(elmtl,ncl)
        /* ensemble des numéros de cliques contenant elmtl */
      icl := icl  ecl
        /* l'intersection de icl et ecl donne l'ensemble des numéros de
        cliques contenant le squelette de gcp et l'adjectif elmtl */
      si icl = vide
      alors
        listeadject := DELETEMT(listeadject,elmtl)
          /* dans ce cas supprimer elmtl de listeadject */
      sinon
        resteadject := DELETEMT(listeadject,elmtl)
          /* restadject a pour valeur listeadject épurée de elmtl */
        elmt2 := PREM(resteadject) /* premier élément de resteadject */
```

```
tantque resteadject ≠ vide
faire
  icll := LIRE(elmt2,ncl)
    /* icll est l'ensemble des numéros de cliques contenant
    elmt2 */
  icll := icl  icll
    /* icll est l'ensemble des numéros de cliques contenant le
    squelette de gcp, elmtl, et elmt2 */
  si icll ≠ vide
  alors
    adjectinter := adjectinter U elmt2
      /* adjectinter contient le plus grand sous-ensemble
      d'adjectifs donnant un "pattern-matching" réussi, et
      contenant elmtl */
    icl := icll
  finsi
  elmt2 := SUIV(elmt2) /* élément suivant dans elmt2 */
finfaire
si CARD(adject.groupe) < CARD(adjectinter)
alors
/* adject.groupe a une cardinalité minimale par rapport à
adjectinter, il est remplacé par adjectinter */
  adject.groupe := adjectinter
  adject.numcl := PREM(icl)
finsi
si CARD(adject.groupe) >= CARD (listeadject) - 1
alors
  trouvé = vrai
    /* le sous-ensemble d'adjectifs formé diffère de 1 de la
    liste d'adjectifs initiale ; le traitement est fini et le
    résultat est dans adject */
finsi
finsi
elmtl := SUIV(listeadject)
  /* élément suivant dans listeadject */
finfaire
----> adject
FIN
```

```
FONCTION RESTAT (gcp : TGCP) ----> EGI
    /* pour un gcp initial RESTAT renvoie l'ensemble d'éléments de type
    TGI qui découlent de la décomposition du gcp via le réseau
    syntaxique */
DEBUT
    fils := FILSMAX(gcp,MODELE(gcp))
      /* FILSMAX renvoie les deux sous-structures maximales éventuelles
      qu'on peut extraire du gcp via le réseau statique */
    si fils = vide
    alors
      result := vide
    sinon
      elmt := PREM(fils)   /* premier élément de l'ensemble des fils */
      tantque fils ≠ vide
      faire
        result := result U PATTERN(elmt)
        elmt := SUIV(fils)  /* élément suivant dans fils */
      finfaire
      result := EPUR(result)
        /* EPUR renvoie result amputé des occurrences multiples de ses
        éléments */
    finsi
    /* si gcp n'est pas découpable, result est à vide; dans le cas
    contraire, result contient l'ensemble des gcf résultat de PATTERN
    avec la ou les deux sous-structures maximales du gcp initial */
    ----> result
FIN
```

```
FONCTION FILSMAX (gcp : TGCP, modèle : TCAR) ----> EGCP
    /* pour un gcp ayant pour modèle modèle, FILSMAX renvoie les deux
    instanciations éventuelles des sous-structures maximales extraites
    pour le modèle, via le réseau syntaxique */
DEBUT
    modepur := DELETEMT(´a,modèle)
      /* modepur prend pour valeur l´ensemble modèle épuré de la
      catégorie adjectif */
    structmax := LIRE(rs,modepur)
      /* structmax a pour valeur éventuellement, la ou les deux
      sous-structures maximales de modepur extraites du réseau */
    si structmax ≠ vide
    alors
      élement := PREM(structmax)   /* premier élément de structmax */
      tantque structmax ≠ vide
      faire
        result := result U COMP(gcp,élément)
        /* COMP renvoie l´instanciation de élément avec gcp */
        élément := SUIV(structmax) /* élément suivant dans structmax */
      finfaire
    finsi
    /* si le modèle n´est pas décomposable result est vide, sinon
    result est l´union des instanciations respectives des
    sous-structures maximales du gcp */
    ----> result
FIN
```

```
FONCTION INDEXATION (lgcf : TGCF; tnum : ENTIER; poids : REEl)
 ----> TSSCL
    /* pour chaque élément lgcf, contenant un numéro de clique et un
    GCF, INDEXATION renvoie la valeur de la relation d'indexation créée
    ou mise à jour dans la clique entre GCF et l'unité d'indexation de
    numéro tnum (poids est à 1 si les éléments de lgcf ont été extraits
    d'un titre) */
DEBUT
    clique := lgcf.numclique  /* le numéro de clique à mettre à jour */
    gcf := lgcf.groupe        /* le GCF à considérer */
    ssclique := PARCOURS(gcf,clique)
      /* si dans numclique il existe une sous-clique contenant gcf,
      PARCOURS renvoie cette sous-clique, sinon vide */
    si ssclique = vide
    alors
      ----> CREESSCL(clique,gcf,tnum,poids)
        /* CREESSCL renvoie la sous-clique créée qui contient la
        relation d'indexation entre gcf et tnum */
    sinon
      ----> MISJSSCL(ssclique,tnum,poids)
        /* MISJSSCL renvoie la sous-clique mise à jour avec la relation
        d'indexation entre gcf et tnum */
    finsi
FIN
```

```
FONCTION CRESSCL (numclique : ENTIER; groupe : TGCP; tnum : ENTIER;
poids : REEL) ----> TSSCL
    /* CRESSCL renvoie la sous-clique créée dans la clique de numéro
    numclique contenant la relation d'indexation entre groupe et le
    texte tnum */
DEBUT
    ssclique.groupe := groupe     /* le GCF considéré */
    ssclique.frectot := 1       /* la fréquence totale du GCF */
    elmt := 1
    avec ssclique.relation[elemt]
    faire
      frecloc := 1  /* fréquence locale du GCF */
      si poids = 1
      alors
        prel := 1
          /* si le GCF appartient à un titre, son poids d'indexation
          est mis automatiquement à 1 */
      sinon
        prel := 0
          /* sinon il est à 0 jusqu'à sa réévaluation à la fin du
          traitement de l'entité */
      finsi
      ref := tnum /* référence à l'unité d'indexation traitée */
    finavec
    clique := PLACECL(numclique,ssclique)
      /* concatène à la clique numéro numclique la sous-clique créée */
    ----> ssclique
FIN
```

```
FONCTION REPERFRECTOT (groupe : TGCP; tnum : ENTIER) ----> ESSCL
    /* REPERFRECTOT renvoie l'ensemble des sous-cliques à mettre à jour
    et contenant les GCF éventuels restreints extraits de groupe ; tnum
    est le numéro de l'unité d'indexation traitée */
DEBUT
    descendant := EXTRFILS(groupe)
      /* EXTRFILS(groupe) renvoie l'ensemble des GCF, précédés des
      numéros de cliques, qu'on peut extraire de groupe */
    si descendant = vide
    alors
      result = vide        /* le groupe ne donne pas de GCF */
    sinon
      elmt := PREM(descendant)  /* premier élément de descendant */
      tantque descendant ≠ vide
        ssclique := MISJOURFREQ(élément,texte)
          /* MISJOURFREQ(élément,texte) renvoie la sous-clique
          éventuelle qui a été mise à jour */
        si ssclique ≠ vide
        alors
          result := result + ssclique
        finsi
        elmt := SUIV(descendant)  /* élément suivant dans descendant */
      finfaire
    finsi
    ----> result
FIN
```

```
FONCTION EXTRFILS(groupe : TGCP) ----> EGI
    /* pour un groupe ayant la structure de GCP, EXTRFILS renvoie
    l'ensemble des GCF plus petits que le groupe pour lesquels on doit
    mettre à jour les fréquences */
DEBUT
    modèle := MOD(OSSATURE(groupe))
      /* MOD renvoie le modèle syntaxique épuré de ses adjectifs */
    choix
      -modèle = vide : result = vide
      -modèle = 's
        choix
         -CARD(groupe) = 1 : result := vide
           /* le groupe est un substantif, il ne donne pas de
           décomposition */
         -CARD(groupe) = 2 : result := INTER(OSSATURE(groupe))
           /* le groupe contient un substantif suivi de son adjectif,
           il se décompose en substantif */
         -CARD(groupe) > 2 : result := JEUXADJECT(groupe)
           /* JEUXADJECT(groupe) renvoie l'ensemble des GCF que l'on
           peut extraire de groupe en considérant uniquement les
           adjectifs */
        finchoix

      -modèle ≠ vide et modèle ≠ 's :
        fils := FILSMAX(groupe)
          /* FILSMAX renvoie l'ensemble éventuel des deux
          sous-structures maximales de groupe via le réseau syntaxique
          */
        choix
          -fils = vide :        /* le groupe est indécomposable */
            result := JEUXADJECT(groupe)
             /* les GCF restreints éventuels sont obtenus uniquement
             avec le jeux d'adjectifs */
          -fils ≠ vide et modèle ≠ 'SPS :
            result := JEUXADJECT(groupe) U EXTRFILS(PREM(fils))
                      U EXTRFILS(SECOND(fils))
```

```
/* si le groupe est complexe, result est l'union, de
l'ensemble des GCF qu'on peut extraire avec le jeux
d'adjectifs, ainsi que les deux ensembles de GCF extraits
respectivement des deux sous-structures maximales de
groupe */
-fils ≠ vide et modèle = 'SPS :
  prep := PREP(groupe)
    /* le modèle de groupe est 'SPS et prep est la
    préposition qu'il contient */
  list1 := JEUXADJECT(PREM(fils--
    /* list1 est l'ensemble des GCF qu'on peut extraire avec
    le jeu d'adjectifs de la première sous-structure maximale
    de groupe */
  list2 := JEUXADJECT(SECOND(fils))
    /* list2 est équivalent à list1 pour la seconde
    sous-structure maximale de groupe */
  result := PLUSOC(list1 U list2)
    /* un sous-ensemble de result est l'union de list1 et
    li t2 */
  elmt1 := PREM(list1)
    /* premier élément de list1 */
  tantque list1 ≠ vide
  faire
    elmt2 := PREM(list2)  /* premier élément de list2 */
    tantque list2 ≠ vide
    faire
      élément := FORMGROUPE(elmt1,prep,eltm2)
      /* FORMGROUPE  renvoie un couple formé d'un numéro de
      clique et d'un GCF formé à partir de elmt1, prep, elmt2
      */
      si élément ≠ vide ET NON(élément dans result)
      alors
        result := result U élément
        /* si élément n'existe pas dans result, on le
        concatène à result */
      finsi
```

```
                elmt2 := SUIV(list2)    /* élément suivant dans list2 */
              finfaire
              elmtl := SUIV(listl)    /* élément suivant de listl */
            finfaire
          finchoix
        finchoix
          ----> result
FIN
```

```
                ,YNAM (entit : ENT) ----> EGI
                une unité non  minimale entit, INDEXDYNAM renvoie  le sous
        ..mble  de termes  d'indexation  communs aux  filles directes  de
    entit */
DEBUT
    si entit > nivmin
      /* nivmin est le niveau de l'unité d'indexation minimale */
    alors
      listfille := FILLE(entit)
        /* listfille contient l'ensemble des noeuds fils de entit */
      remonterm := INTER(entit,listfille)
        /* remonterm contient le  sous ensemble des termes d'indexation
        communs aus éléments de listfille */
      élément := PREM(remonterm)
      tantque remonterm ≠ vide
      faire
        CREREL(élément,entit)
          /*  CREREL  value  et  crée la  relation  d'indexation  entre
          élément et le niveau entit */
        élément := SUIV(remonterm)
      fintantque
    finsi
    ----> remonterm
FIN
```

AUTORISATION de SOUTENANCE

VU les dispositions de l'article 3 de l'arrêté du 16 avril 1974,

VU le rapport de présentation de Monsieur le Professeur Y. CHIARAMELLA

Mademoiselle KERKOUBA Dalila

est autorisée à présenter une thèse en soutenance en vue de l'obtention du titre de DOCTEUR de TROISIEME CYCLE, spécialité "Informatique".

Fait à Grenoble, le 6 novembre 1984

Le Président de l'I.N.P.-G

D. BLOCH
Président
de l'Institut National Polytechnique
de Grenoble

P.O. le Vice-Président,