Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'information : de l'écrit vers la gestion des connaissances

Sahbi Sidhom

n° d'ordre : 29-2002

THÈSE

présentée devant

L'UNIVERSITÉ CLAUDE BERNARD – LYON 1
Ecole Doctorale Informatique et Information pour la Société

pour obtenir le titre de

DOCTEUR DE L'UNIVERSITÉ

spécialité
Informatique & Sciences de l'Information et de la Communication

par

Sahbi SIDHOM

Sujet de la thèse :

Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'information :

de l'écrit vers la gestion des connaissances

Sous la Direction de : Pr. Mohamed HASSOUN

soutenue le : Lundi 11 Mars 2002

Jury :

M. Richard BOUCHÉ	(Examinateur)	– ENSSIB Lyon (Equipe SII)
M. Amos DAVID	(Rapporteur)	– Université de Nancy 2 (INRIA-Lorraine)
M. Mohamed HASSOUN	(Directeur de Thèse)	– ENSSIB Lyon (Equipe SII)
M. Jean-Marie PINON	(Examinateur)	– INSA de Lyon (Laboratoire LISI)
Mme Violaine PRINCE	(Rapporteur)	– Université de Montpellier 2 (LIRMM-CNRS)
M. Jérôme VINESSE	(Examinateur)	– France Télécom R&D Lannion (URD LangNat)

À La Mémoire de

mon père Hassen SIDHOM,

je dédie cette thèse

Avant-propos

Ce document est rédigé au laboratoire des *Systèmes d'information et Interfaces* (SII) à l'*Ecole Nationale des Sciences de l'Information et des Bibliothèques* (ENSSIB) à Lyon.
Le travail présent est le résultat d'une activité de recherche s'échelonnant sur plusieurs années. Il a été effectué dans le cadre du Programme Doctoral en Sciences de l'Information et de la Communication de l'Ecole Doctorale Lyonnaise : Informatique et Information pour la Société (EDIIS). L'équipe SII travaille sur l'étude des systèmes d'information de type documentaire et de leurs interfaces y compris dans leur dimension multilingue.

Concernant le groupe de recherche SYDO (Systèmes Documentaires), à son origine, il rassemblait des chercheurs des universités de Lyon, Grenoble (France) et Fribourg (Suisse). Au départ, les membres de ce groupe animés par les professeurs A. Berrendonner, R. Bouché, M. Le Guern et J. Rouault se sont réunis très régulièrement entre 1979 et 1986 en coordonnant des travaux de recherches sur l'analyse documentaire orientées vers l'analyse linguistique automatique du français. Leurs travaux de recherche sont considérés comme l'édifice théorique sur la conception des systèmes d'indexation documentaires.

Plus tard, le groupe a continué d'être animé par les professeurs R. Bouché, M. Le Guern et tant d'autres chercheurs à Lyon, au sein du laboratoire d'informatique documentaire à l'université Lyon1, puis au Centre d'Etudes et de Recherches en Science de l'Information (CERSI) regroupant les établissements universitaires ENSSIB, Lyon 1 et Lyon 2. Actuellement, quelques membres du groupe continuent leurs activités au sein du Laboratoire SII à l'ENSSIB.

La genèse des travaux de recherche SYDO a longtemps fait porter ses efforts sur le développement d'outils statistiques et lexicomatiques. En développant parallèlement des recherches et des projets sur l'analyse morphologique et syntaxique en langue naturelle, le groupe s'est résolument orienté vers les fondements théoriques de l'indexation documentaire. Il propose une approche linguistique et se focalise sur l'analyse automatique du français puis vers d'autres langues naturelles (anglais, allemand, arabe, etc.).

Je voudrais enfin signaler que ce travail s'est imprégné des réflexions théoriques et pratiques lors des collaborations de recherche avec –l'équipe de recherche en Ingénierie des Informations Documentaires et des Objets Complexes du Professeur Jean-Marie Pinon à l'INSA de Lyon, avec des –spécialistes de l'INAthèque et l'INA-actualités à l'Institut National de l'Audiovisuel (INA) à Paris, et avec –l'équipe Langues Naturelles de France Télécom R&D au Technopole Anticipa à Lannion.

Remerciements

Nous tenons à remercier les responsables du laboratoire des *Systèmes d'information et Interfaces,* qui m'ont offert constamment le soutien et les encouragements sans lesquels ce travail n'aurait pas vu le jour.

Mes remerciements vont à mon Directeur de thèse Mohamed Hassoun qui m'a orienté et conseillé tout au long de mes activités de recherche en thèse. J'ai pu ainsi bénéficier de ses connaissances et de ses compétences dans de très nombreux domaines, en particulier dans le domaine du traitement automatique des langues naturelles, de son art d'orienter et de diriger sans froisser.

Mes remerciements vont aussi, à Richard Bouché, avec lequel j'ai collaboré au sein de son projet DEBORA® (Digital accEss to BOoks of the RenAissance), et qui a été financé par l'Union européenne (DG.xiii). Son objectif ayant été de développer des outils d'accès distant à des collections numérisées du XVIe siècle dans le but d'élargir le public concerné par les collections patrimoniales. R. Bouché est un des précurseurs, en France, de la discipline « Sciences de l'Information et de la Communication », contribue activement à son développement tant au niveau de l'enseignement qu'à celui de la recherche. Par son caractère ouvert, initiateur et de ses connaissances avec lequel j'ai toujours eu des échanges fructueux.

Mes remerciements vont aussi à Jean-Marie Pinon qui m'a encadré et qui m'a fait bénéficié de ses connaissances et celles de son équipe au début de ma thèse dans de nombreux domaines, de son art d'orienter et de ses précieux conseils qu'il m'a apportés.

Mes remerciements vont également à Jérôme Vinesse et ses collaborateurs dans l'équipe Langues Naturelles – France Télécom R&D à Lannion pour m'avoir fait bénéficier d'une activité en recherche et développement sur le projet ERUDIT® : Prototype pour la création de services de renseignements en langages naturels.

Mes remerciement s'adressent également aux membres du jury de thèse, à Violaine Prince et Amos David, pour l'honneur et l'intérêt qu'ils me font en acceptant de lire et de juger ce travail.

Je remercie aussi Colette Lustière, Daniel Dégez-Vataire, Christine Barbier Bouvet, pour les quelques réunions de travail à Paris qu'elles aient été faites dans le cadre de la collaboration avec l'INA, de leurs aides précieuses pour la constitution du corpus de travail sur l'audiovisuel et de m'avoir fait partager leurs expériences professionnelles sur l'étude du corpus.

Mes remerciements,

- à Michel Le Guern pour ses précieux conseils et encouragements dans le domaine,

– à André Deweze et Marcilio De Brito pour leurs conseils et les expériences échangées lors des réunions de travail à l'ENSSIB,

– à tous les membres, passés et présents du laboratoire des *Systèmes d'information et Interfaces* à l'ENSSIB de Lyon avec lesquels j'ai eu des échanges fructueux, et en particulier, à Samuel Tiétsé,

– à tous les membres du laboratoire de recherche en Ingénierie des Informations Documentaires et des Objets Complexes au LISI à l'INSA de Lyon,

– à mes collègues en recherche et développement au Département Communication Homme-Machine à l'IRSIT de Tunis,

– à mes collègues enseignants universitaires à l'Institut de la Communication de l'université Lumière Lyon 2 dont les encouragements, exprimés de diverses manières, ont été fort utiles à la clôture de ce travail.

Je saisie cette opportunité pour remercier :

– Mes familles et mes amis m'ont épaulé tout au long de cette thèse. Je les remercie du fond du coeur. Merci à ma maman, à mes frères et tout particulièrement au très cher frère Selim dont le soutien et la patience m'ont permis de garder le sourire dans les moments difficiles.

SOMMAIRE

Introduction

La diversité des applications réunies de nos jours sous le terme « industrie de la langue » recouvre plusieurs pistes de réflexion. Notre travail de recherche a consisté à baliser le terrain de ce que l'on convient d'appeler « traitement automatique de la langue naturelle ». Pourtant, l'analyse linguistique automatique se trouve au confluent de plusieurs disciplines que sont : la linguistique, la psycholinguistique, l'informatique et les mathématiques.

La démarche scientifique dans ce cadre de travail pluridisciplinaire, nous autorise à prendre objectivement connaissance des problématiques épistémologiques dans chaque cadre d'objet d'étude. Nous nous servirons des concepts, des préceptes et des méthodes relativement concurrents pour aborder ce travail dans chacune de ces disciplines.

En terme de choix et de qualité nous avons proposé une intégration de chaque objet d'étude et ses interactions (relations) avec les autres. En terme d'interaction linguistique-informatique, la présentation de la méthode nous permettra d'allier élégamment les concepts de chaque objet. Nous suggérons ensuite la construction de la structure syntaxique détaillant successivement la technique d'analyse des syntagmes et leur mise en relation.

Au début de notre étude, nous étions confrontés à un objet type qui est l'écrit comme résultat d'une production intellectuelle humaine. La forme de l'écrit est d'une grande variabilité, car elle est soumise à plusieurs facteurs extra-linguistiques qui affectent aussi bien les conditions de production que leurs auteurs. La genèse même de l'écrit, à travers des études anthropologiques, a été soumise aux mêmes types de contraintes.

L'écrit n'en reste pas moins observable à travers notre analyse. Cette caractéristique fondamentale va nous permettre de dresser les structures syntaxiques types à travers les textes étudiés. Cela nous permettra de construire des outils dans le but d'explorer et de formuler des hypothèses sur les structures textuelles, puis de confronter nos hypothèses à la réalité de l'objet lui-même : l'analyse de l'écrit.

Dans cette première démarche de l'étude, il ne s'agit pas d'un objet immatériel comme la *compétence* d'écrire, la disposition d'un sujet à travers ses *connaissances* de la langue, ou la capacité de *formalisation* de l'objet d'étude. Mais bien au contraire d'une production réelle, concrète et issue de la performance, c'est à dire la mise effective du langage à travers l'acte de l'écriture.

Concrètement, les écrits prennent la forme d'une collection de textes. Les textes de l'étude ne sont pas soumis à des contraintes rédactionnelles, de styles ou de profit de sélection aux auteurs. Le recours à cette liberté d'écriture, par extension, va permettre un meilleur aperçu de la couverture des structures conjecturées et par ailleurs leur validation dans la pratique par l'analyse automatique.

Bien que le choix de nos textes soit particulier, car seulement l'élaboration du contenu est en fonction d'une grille d'analyse, cette contrainte de l'*Analyse Documentaire* préserve la variabilité de la forme des productions de textes. Mais, tous restent sous la même influence qu'est l'activité de l'écriture. C'est cette activité qui nous a conduit à considérer le texte dans sa globalité et la phrase ou les structures sous-jacentes dans leur particularité, comme cadre d'étude pertinent.

Un individu plongé dans une activité d'écriture préconise au moins l'existence d'une idée mentale ou non sur un sujet. Ce type d'énoncé nous offre la matière première à l'étude des structures syntaxiques comme objet d'étude linguistique. Ainsi, le texte est vu comme un ensemble cohérent d'unités plus ou moins complexes. Chaque unité s'articule avec les autres et contribue à la réalisation d'un équilibre structurel.

C'est précisément dans ce cadre général de construction de textes qu'il nous faudra réaliser l'analyse syntaxique et aboutir à une interprétation significative : analyse textuelle et recherche d'information.

Dans la démarche imposée lors de l'analyse des différentes structures de notre objet (texte écrit), nous avons cherché en syntaxe la modélisation qui permet de retrouver au mieux les phénomènes linguistiques. Mais également, pour définir une structure de surface, identifier la projection de l'ordre structural suivant un ordre linéaire : emboîtement de certains types de structures syntaxiques.

Dans cet esprit, les travaux du Groupe de Recherche (SYDO : SYstème DOcumentaire), portant sur la découverte des structures formelles de la langue (cas du français, puis étendu à d'autres langues naturelles), se sont révélés essentiels et ils ont contribué à la validation de nos hypothèses. C'est ainsi que la notion de syntagme nominal comme unité pivot dans la structure syntagmatique de la phrase se trouve consolidée, d'une part, par la mise en évidence de sa stabilité, et d'autre part, par la relation de référence, c'est à dire la relation entre les mots et les choses.

Il est commun dans notre domaine de s'offrir la liberté de voir l'écrit comme objet qui a évolué avant de se stabiliser sous sa forme actuelle. La genèse de l'écrit a duré plusieurs milliers d'années avant et après notre ère. Les études anthropologiques ont montré que l'écrit à son origine était comme un objet multidimensionnel véhiculant plusieurs messages. Sa présentation comme sa nature a évolué d'un objet graphique complexe à un objet symbolique linéaire.

La recomposition de l'objet écriture vers une forme stable offre un champ étendu d'interprétation. Le codage de la structure, qui une fois repérée et analysée, va nous permettre de tendre vers un de nos objectifs, à savoir l'analyse morpho-syntaxique automatique.

Aujourd'hui, on compte parmi les mots structurants dans les syntagmes, dans les phrases et dans les paragraphes, des éléments tels que les nominaux, les verbes, les adverbes, les prépositions, les conjonctions, etc., et les accords en genre, en nombre, et en personne. D'autres, ayant un statut particulier comme la ponctuation affectent la forme des énoncés et leur étude.

Dans cette perspective d'étude et le choix porté à un modèle morpho-syntaxique particulier, il ne faut pas en effet perdre de vue que la qualité des résultats d'un analyseur morpho-

syntaxique placé dans un système de traitement de la langue naturelle puise ses performances, d'une part, de la qualité de sa conception (formalisme d'implémentation), et d'autre part, de la qualité des recherches menées en syntaxe (modèle théorique). Plus les concepts théoriques et pratiques seront en accord avec la nature de l'objet d'étude, et meilleures seront la qualité et l'efficacité des résultats de l'analyseur.

N'oublions pas que l'analyseur morpho-syntaxique en question est un module-clé dans toute application du traitement de la langue naturelle. Après confrontation des possibilités théoriques et l'adéquation des outils pratiques à réaliser, un second objectif consiste à produire l'analyseur utilisable dans un tel système et environnement. Un tel analyseur ne sera qu'un produit dérivé et une retombée opératoire de notre recherche première, en linguistique computationnelle, en syntaxe *versus* l'indexation, sur corpus de l'audiovisuel INA France.

Le corpus utilisé pour l'observation et la manipulation est constitué de textes de la langue française. Une hypothèse faite sur sa nature était de type texte libre et sans contraintes de style de rédaction. La variété de son contenu décrit sa représentativité comme document secondaire à celui qu'il représente : texte résumé sur le document primaire, c'est à dire image (fixe ou animée), son (séquence audio), vidéo, et texte. Notre source à l'origine était constituée des bases de données documentaires de l'audiovisuel : INAthèque et INA-actualités à Paris.

En confrontation de l'étude menée sur le corpus (les résumés de contenu INA) et l'étude linguistique (basée sur l'extraction des syntagmes nominaux), des régularités structurelles et syntaxiques ont été observées dans les textes résumés. Cette source de régularité, dans ce travail, était la base de la construction de la grammaire de réécriture de l'analyseur.

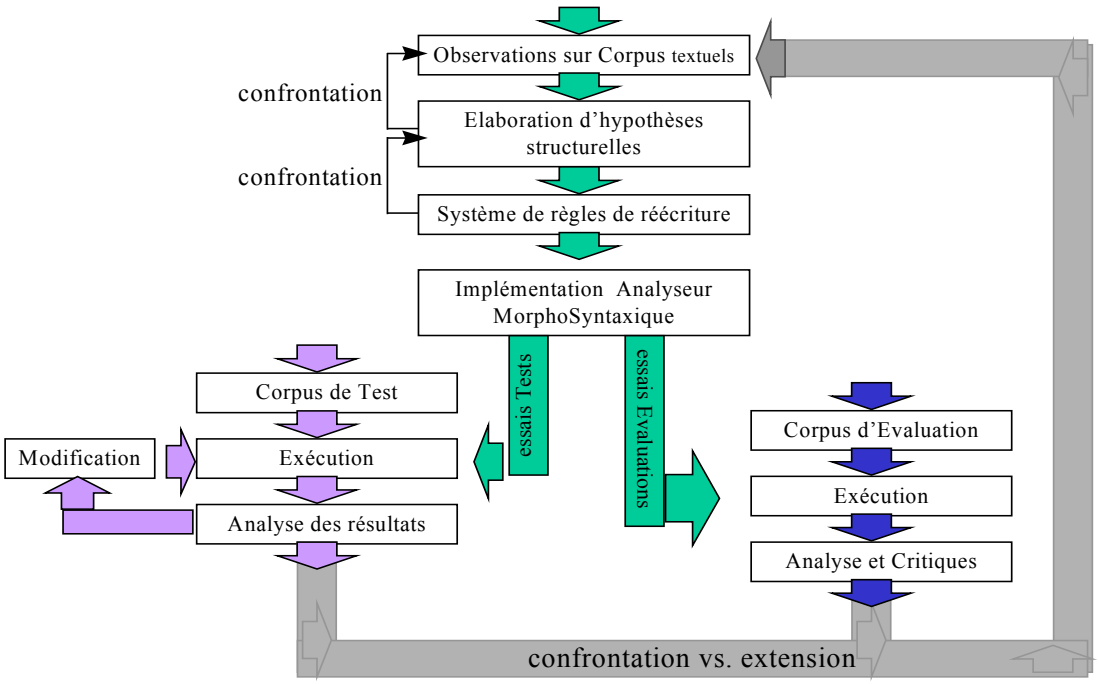


Fig.1. Démarche structuraliste sur les objets d'étude.

De manière explicite, la démarche scientifique (*Fig.1.*) que nous avons suivie pour étayer les hypothèses de notre travail et corréler nos choix théoriques avec nos conceptions pratiques, consiste à :

(i) l'élaboration d'hypothèses sur les structures syntaxiques, qui se concrétise par l'étude linguistique sur le corpus de textes résumés,
(ii) la transcription des observations faites sur corpus en système stable de règles de réécriture grammaticale,
(iii) la matérialisation du système de réécriture par l'implémentation de l'analyseur morpho-syntaxique, et
(iv) l'évaluation de l'analyseur par application directe sur corpus et sa comparaison aux observations retenues dans les hypothèses de l'étape (i) et la couverture grammaticale de l'étape (ii).

Au début de cette étude, le travail sur corpus pour construire des hypothèses structurelles à travers les observations effectuées est un travail de longue haleine. Leurs confrontations à l'évaluation entre la grammaire de réécriture avec le corpus se sont réalisées. Ainsi, l'appréciation résultante du système sera toujours un compromis entre les connaissances acquises sur corpus pour les soumettre à l'évaluation dans une phase opératoire du système. Il s'agira dans ce cas d'une recherche de l'optimisation qualitative sur les résultats et d'autre part, l'extension des capacités du système pour la prise en compte de nouveaux phénomènes (nouvelles hypothèses) sur de nouveaux corpus : recherche de la robustesse quantitative sur les connaissances du système.

Cette démarche dispose de pleins de cohérences et de précisions sur l'aptitude à conserver une indépendance entre les variétés de corpus (corpus de tests, d'évaluation, etc.) et la robustesse du système vis-à-vis du traitement de nouveaux phénomènes. Les réalisations se révèlent plus contraignantes, car les attendus en question réclament en effet un travail manuel et intellectuel très important pour offrir des résultats concis.

L'objectif de ce travail est de construire *une Plate-forme d'analyse morpho-syntaxique* pour l'indexation automatique et la recherche d'information. Elle est composée d'un –noyau d'indexation automatique (*processus d'indexation*) qui utilise le modèle des syntagmes nominaux comme descripteurs de l'information textuelle. Ces travaux s'orientent vers –l'organisation des syntagmes nominaux (hiérarchie et emboîtement) pour la recherche d'information (*processus de recherche*) dans une base de données textuelles. Cette Plate-forme d'analyse dans sa logique de fonctionnement sera un outil d'investigation orienté vers l'organisation et la gestion des connaissances écrites.

Nous débutons ce mémoire, chapitre I, par une étude sur la connaissance et la transmission des savoirs.
Dans le contexte de l'analyse des contenus audiovisuels, la recherche d'un lien commun entre image, parole et texte n'était pas évident. En remontant dans le temps, des recherches sur les origines de *l'activité langagière* de l'homme ont, en particulier, montré que l'écriture avait ses racines dans les arts graphiques et dans des dessins signifiants (des indices figuratifs et d'autres arbitraires de type abstrait). Plus tard, l'écriture de plus en plus abstraite est devenue un vecteur de correspondance qui fait le lien entre culture visuelle, culture orale et culture écrite.
Le chapitre II, sur l'état de l'art dans le domaine de l'indexation automatique, recouvre les activités documentaires en matière d'analyse de contenu et sa dynamique évolutive (intellectuelle, semi-automatique et automatique).
Nous évoquons dans cette partie les problèmes rencontrés dans chaque méthode ou approche étudiées. Certes, les solutions en matière de recherches sur l'indexation automatique ne sont

pas définitives, mais les approches semblent résoudre certains problèmes essentiellement liés aux traitements du langage naturel.

Le modèle linguistique du Groupe de recherche SYDO, objet du chapitre III, est le thème central dans la 're-formulation' de la problématique classique du descripteur en indexation documentaire. Nous y exposerons un modèle linguistique qui rend compte de la problématique posée en général par les systèmes d'information documentaire et en particulier le statut d'un descripteur en indexation automatique.

Jusqu'ici, notre contribution a consisté à définir notre problématique et le contexte de travail où nous nous situons avec ses complexités. Nous utilisons également des espaces de références pour le situer afin d'inscrire nos hypothèses méthodologiques dans le but d'accomplir des tâches de recherche et développement.

Le chapitre IV exposera l'Extraction des Connaissances dans les processus d'analyse et d'indexation du contenu (audiovisuel) basée sur corpus.
Génération de résumés, création d'outils linguistiques automatiques, mémorisation des connaissances et recherche d'information, sont toutes des activités *'intelligentes'* où nous accordons une place très importante aux traitements automatisés ou semi-automatisés. Dans ce cadre précis, le thème de discours se matérialise sous une forme linguistique spécifique : le syntagme nominal.
L'approche logique du syntagme nominal a montré en effet que l'identité de l'objet descripteur était une construction d'un parcours interprétatif des usagers à travers l'espace documentaire.

Le chapitre V décrira une méthode de conception du noyau d'indexation automatique : la réalisation des différents outils automatiques servant à l'analyse du langage naturel et l'implémentation de l'analyseur morpho-syntaxique pour le français écrit. L'architecture de l'application est basée sur les automates à transitions augmentées en cascade (CATN) de W. Woods (1970-80).

Au chapitre VI, nous naviguerons autour de l'objet 'descripteur' construit comme un syntagme nominal et relevant d'un principe d'organisation du discours. Il importe sur ce point de pouvoir distinguer la phase d'extraction des unités de discours (l'indexation), la phase de réorganisation de ces unités et leur classification (la gestion des connaissances). Cette dernière étape permet l'exploration des sources de connaissances selon les constructions logico-sémantiques des syntagmes nominaux.

Dans la conclusion, nous présentons les limites du développement actuel concernant la plate-forme d'indexation automatique et de recherche d'information ainsi que les perspectives de cette recherche.
Dans la partie annexe, nous complétons ce travail concernant la *plate-forme* par les interfaces graphiques et leurs modes d'exploitations. Ces interfaces sont l'aboutissement de la conception entreprise, aux choix effectués et aux implémentations réalisées puis inter-connectées, des différentes tâches du projet. Le premier processus de construction des interfaces Homme-Machine concerne les ressources linguistiques : dictionnaire électronique du français. Le second processus concerne l'analyse morpho-syntaxique (noyau d'indexation). En dernier, le processus concerne la présentation de l'organisation des connaissances autour du syntagme nominal et offre des scénarii différents de recherche d'informations.

CHAPITRE I :

La connaissance et la transmission des savoirs : du savoir visuel au savoir écrit

I. Introduction

L'histoire de la connaissance et de la transmission des savoirs a été marquée par deux innovations majeures : – l'invention de l'écriture, et – la « généralisation » de la forme scolaire par l'apprentissage, plus tard, du manuscrit à l'invention de l'imprimerie, donnant accès à l'écrit.

De nos jours, l'élaboration de l'écrit sous toutes ses formes a donné naissance à une prolifération croissante des sources documentaires. Cette capacité de production des connaissances a permis aux hommes de retrouver le terrain favorable pour développer une attitude rationnelle d'inspection de ces connaissances, de les étudier, de les commenter et de les faire évoluer.

A travers les temps, la pensée logique et l'attitude rationnelle de l'homme se sont construites peu à peu et se sont complétées par accumulation des sources de connaissances. L'origine de ces sources, à travers notre étude basée sur les travaux de l'anthropologue Jack Goody, n'est pas aussi évidente. Sans doute que la base physique de l'écriture soit clairement la même que celle du dessin, de la gravure et de la peinture : les arts graphiques. Si l'écriture à donc ses racines dans les arts graphiques, dans des dessins signifiants, le niveau d'abstraction de l'écriture pour représenter le monde est loin de rester graphique [ALLOTT, 2001], [ALLOUCHE, 99].

Le motif graphique penche vers le pôle pictural ou vers le pôle arbitraire ou formel, sa forme influe sur le rapport entre le signifiant et le signifié. En d'autres termes, le dessin graphique à son origine était une mesure profonde du langage enfoui dans l'activité humaine. Comme soutenant de cette hypothèse A. Leroi-Gourhan, J. Goody et tant d'autres chercheurs [GOODY, 94] : *l'art figuratif est inséparable du langage*.

Il nous a semblé évident, dans le contexte de l'analyse des contenus audiovisuels ou multimédia et particulièrement la recherche d'un lien commun entre image, texte et parole dans le contexte d'application documentaire, de remonter dans le temps et de rechercher les origines de *l'activité langagière* de l'homme (sous toutes ses formes). Surtout que cette activité avec l'avènement actuel de l'Internet s'est d'autant développée et proliférée sans limites [ANDRIEU, 96a-b], [BOURSIER, 94], [LEVY, 95], [LIPPMAN, 96].

De nos jours, la prolifération du multimédia et des hypermédias en général nous impose une activité d'analyse et de verbalisation du contenu [MOULIN, 91], [MARCHIONINI, 95a-b]. Cette tache n'est pas évidente car elle requiert l'analyse du langage graphique, la parole et de l'écrit [MICHEL, 96], [VIEIRA, 96], [CASANOVA, 99], [VETTRAINO, 99]. Si l'homme d'aujourd'hui a acquis l'écriture abstraite pour communiquer, pour développer et pour enrichir

son savoir de manière quasi universelle, « l'écriture » graphique lui demeure un obstacle alors qu'elle était acquise par ses ancêtres depuis plus de 6000 ans.

I.1. L'écriture : invention ou révolution ?

Depuis près de 6000 ans, l'invention de l'écriture dans la société de Mésopotamie et de Chine a singulièrement modifié l'histoire de la connaissance et de la transmission des savoirs. Pour l'anthropologue Jack Goody, l'écriture et surtout l'écriture alphabétique, aurait permis aux hommes d'analyser leur propre discours grâce à la forme continue qu'elle donnait au message qu'il soit oral ou écrit. Grâce à cette capacité d'inspection du discours, le champ de l'activité critique (les oppositions entre les modes de pensée et de discours) s'est développé vers la pensée et le discours logique.

Comment se construit le savoir ? et comment s'enrichit-il ?
D'abord, que la différence entre les cultures écrites et les cultures orales n'est pas une question de degré ou de niveau d'intelligence, mais liée à la présence ou l'absence des instruments avec lesquels les intellectuels travaillent. Par ces réflexions « instrumentales », il est facile de percevoir une contradiction entre deux termes employés ou entre deux phrases ou expressions par le simple geste de juxtaposition sur des traces écrites et à tout moment. Il est beaucoup difficile d'en établir ce genre de perception dans une conversation et surtout plus tard après une durée de temps. Comme disait J. Goody [GOODY, 98] :

> « *L'écriture représente une autre façon de mesurer les faits, et la précision de la mesure est un aspect important des systèmes de savoir* ».

L'écriture « alphabétique » comme celle des Grecs n'est plus vue comme une réussite tellement unique, car les écrits *logographiques*, comme le Chinois, contiennent de nombreuses composantes phonétiques qui imposent la séparation entre ses éléments.

En terme de modes de communication, les difficultés liées aux logogrammes lors de l'apprentissage de la lecture et surtout de l'écriture n'entraînent pas un rapport de conflit entre l'écriture et la connaissance. La graphie donne accès d'une façon très différente à la lecture et à la composition d'oeuvres écrites. Cette notion de maîtrise de l'écriture comme indicateur de savoir entraîne le même système d'évaluation dans le cas d'une écriture syllabique ou alphabétique.

I.1.1. Développement historique de l'écriture

Les systèmes de communication sont en rapport direct avec ce que l'homme peut faire de son monde à la fois interne en termes de pensée et externe en termes d'organisations culturelle, sociale et économique. Le langage est l'attribut humain spécifique, - le moyen primordial d'interaction entre les individus et la base du développement de la culture, - la façon dont un comportement enseigné (social ou intellectuel) est transmis de génération en génération, et - le facteur de perception et d'échange des transactions économiques entre individus et communautés.

En terme d'associations qui lient l'attribut humain et son organisation, nous distinguons :

- **Langage et écriture**

  Autant le **langage** est associé à la « *culture* », autant l'**écriture** est liée à la « *civilisation* ». Les conséquences de ces associations ne se résument pas au jeu des

réalisations (comme les objets fabriqués par une moule) ou de pressions exercées (comme pour donner de nouvelles formes aux objets fabriqués) sur l'organisation sociale, mais d'un changement dans le processus cognitif dont l'homme est l'héritier.

- **Ecriture et dessin**

La base physique de l'écriture est clairement la même que celle du dessin, de la gravure et de la peinture (les arts graphiques). La faculté de manipuler des outils au moyen de la main coordonnée avec l'oeil, l'oreille et le cerveau permet à l'homme d'écrire et de dessiner.

L'écriture a donc ses racines dans les arts graphiques, dans des dessins signifiants. De telles activités parviennent depuis les premières phases de l'histoire de l'homme et dont les traces existantes avec l'arrivée de *l'homme Paléolithique supérieur* (30000-10000 avant J.-C.) sous des formes graphiques dans les cavernes, les abris sous roche ou sur les rouleaux en écorce d'arbre. Les conséquences de ces dessins sont soit ***communicatives*** soit ***expressives***.

Les *conséquences expressives* peuvent être vues comme une communication incomplète ou avec soi-même, une sorte de monologue graphique dont le but est l'extériorisation des pensées et des sentiments ou simplement la création du dessin lui-même, sans qu'une communication immédiate intervienne.

Les *conséquences communicatives*, tout en parlant du système d'écriture (Fig.I.1.), montrent que les motifs iconographiques (qualifiés aussi de pictural, de figuratif ou d'eidétique) sont souvent assimilés à des pictographes ou pictogrammes : des signes isolés plutôt que des systèmes étendus d'écriture. Une distinction à l'intérieur du motif signifiant entre le figuratif et l'arbitraire peut correspondre aussi à la distinction sémiotique entre les ***indices naturels*** et les ***signa***. Les indices naturels s'expliquent d'eux-mêmes, mais dans le cas de signa « que A veuille dire B » est le résultat d'un choix humain arbitraire [MULDER, 72], [LEACH, 76].

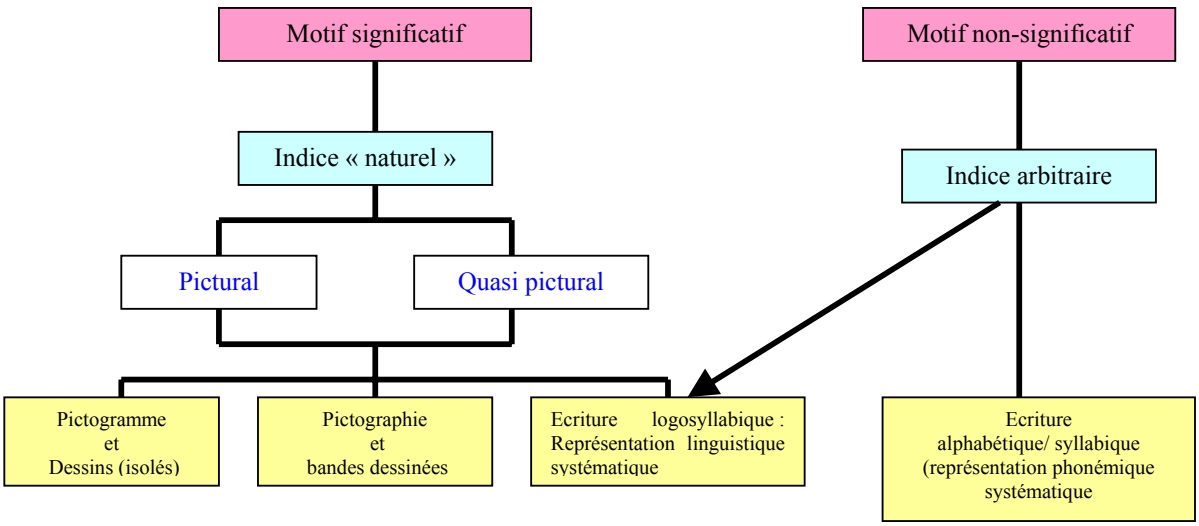


Fig.I.1. Variations dans la représentation graphique [Goody, 94]

Les systèmes graphiques dans leurs débuts (par exemple les pictogrammes) ne sont pas arbitraires car l'association est souvent ou en partie « naturelle ». Un objet (dessiné) peut indiquer une autre sorte d'objet en description graphique et en même temps qu'il indique pour cet objet un morphème ou une description verbale plus étendue.

Cette représentation graphique peut être *figurative* ou *métonymique* (= une partie pour le tout), comme elle peut être de type *stylisé simplifié*. Dans ces cas, les représentations se rapprochent et s'orientent vers les « indices naturels ». Dans tout répertoire graphique, surtout dans un système d'écriture complet, certainement il est question d'introduire des indices arbitraires, c'est-à-dire les *signa*, en sorte que le système ne peut pas être purement figuratif et le langage lui-même n'a pas de lien univoque avec des objets ou des actions du monde extérieur.

Comme l'affirme F. Saussure, le langage est arbitraire, aussi bien cette hypothèse reste valide pour l'écriture graphique : un lien pictographique direct soit possible dans le cas de certains signes logographiques (correspondance à des mots ou à des morphèmes), de même qu'un lien similaire sera possible avec des sons au moyen d'onomatopées dans le cas du langage parlé. Toutefois il faut souligner que même les systèmes précurseurs les plus simples de l'écriture comprennent quelques signes non figuratifs [GOODY, 94].

I.1.2. Proto-systèmes d'écriture

Les signes graphiques dans certaines cultures et chez des types de société ont servi pour apprendre et se rappeler de certains évènements ou évoquer des rituels particuliers comme des chants et des incantations.

Certains anthropologues ont noté l'existence et la correspondance de tel symbole pour tel mot. Ainsi, la pictographie sert de soutien à la mémoire et elle est employée surtout par ceux qui veulent apprendre. Mais dans tous les cas, ces usages étaient marginaux et on peut présumer qu'ils auraient requis la même personne pour les décoder que pour les coder. En fait l'essence même de ces signes paraissent être des aide-mémoires plus qu'un vrai système d'écriture.

Quelle est la nature de cette proto-écriture ?
Souvent appelée par écriture pictographique précédent tous les systèmes d'écriture qui se servent de signes arbitraires : L'élément pictural domine à cause du rapport mnémotechnique entre le signe et le signifié ou plusieurs, car il peut y avoir un nombre considérable d'interprétations possibles d'un signe donné et même d'un texte donné.

Dans les travaux de I.J. Gelb (1952-1974), et reportés par J. Goody, celui-ci distingue dans les signes précurseurs de l'écriture deux classes :

- **Les procédés descriptifs et représentatifs** (appelés *procédés descriptifs*) :
  la compréhension de tels signes ne dépend pas de la connaissance d'une langue particulière. Les procédés descriptifs sont au fond des indices naturels d'espèce statique. Ce type de procédé fut employé par les Indiens d'Amérique pour faire les traités de paix.

- **Les procédés mnémoniques et identifiants** (appelés *procédés mémoriaux*) :
  ils sont employés pour consigner les paroles et non pour décrire. Ils peuvent être abstraits ou figuratifs, et sont des signes de nature séquentielle. Toutefois ce ne sont pas des transcriptions

du langage, mais plutôt une sorte de sténographie figurative : une mnémotechnique qui essaie de rappeler ou de servir d'aide-mémoire à des énoncés linguistiques, plutôt que de les reproduire. Chez les Mycéniens et les premiers Sumériens, l'emploi d'une telle transcription était limité à de simples rapports administratifs.

Dans ces deux procédés, il n'y a pas de lien systématique entre le signe et le son tant que nous ne sommes pas parvenus à de véritables systèmes d'écriture employant des signes représentant des mots (des logogrammes). La nécessité d'un grand nombre de signes semble immédiate pour leur emploi que pour leur transcription. La nature concrète de ces systèmes dépend de sa propre logique interne. Elle n'a aucun trait de la constitution de l'esprit de ses usagers (l'*esprit primitif,* selon J. Goody). Le chinois, qui est le seul grand système logographique encore en usage aujourd'hui, contient précisément un tel élément figuratif, bien qu'une grande partie semble avoir été perdue au cours du temps.

I.2. Premiers systèmes d'écriture

A propos des principes de l'écriture « pictographique », qui sont à la base de l'écriture proprement dite, ne relèvent pas d'une nature totalement différente : actions, personnes et objets ne peuvent être séparés de leurs symboles linguistiques, en sorte que les signes ou les symboles figuratifs opèrent à travers un canal linguistique et visuel.

Le progrès principal, qui a été observé par les scientifiques en la matière, réside dans le niveau auquel le système graphique réussit à doubler (ou amplifier) le système linguistique : la correspondance sémantique entre le mot, le signe et la correspondance phonétique.

Une telle présentation de signes figuratifs était incorporée dans tous les premiers systèmes d'écriture. Mais ils n'étaient pas la seule source de signes pour établir la correspondance entre la langue et la graphie : appelée écriture.

I.2.1. L'écriture logographique

Il est clair que les systèmes logographiques d'écriture se sont développés à partir d'usages plus simples des signes graphiques. Mais l'écriture à part entière qui vint à se réaliser incorpora la représentation systématique et complète des mots et de leurs référents par des signes graphiques particuliers. A l'évidence, beaucoup de mots ont des référents liés au « monde extérieur ».

En prenant l'exemple du signe noté X (au sens de l'écriture), dans sa convention, il signifie « croix », et fait référence au concept, à l'objet, à l'action « croiser » et au son (sens phonétique). Tout en sachant que la référence première de X est liée au son, alors que dans les formules figuratives de la proto-écriture la référence la plus immédiate est liée à l'objet ou à l'événement lui-même.

Chaque type élaboré d'écriture possède certains signes qui représentent des syllabes et des sons phonétiques aussi bien que des mots et donc économisent sur le nombre de signes dont on a besoin. Par exemple, le signe *[an]* ajouté au signe *[faon]* pourrait se lire « enfant ». De ce fait ces premiers systèmes complets sont dits logosyllabiques : ils se servent de signes qui correspondent à la fois à des mots et à des syllabes.

On connaît sept systèmes d'écriture (Tab.I.2.1.) de ce genre dans les anciennes sociétés humaines :

Systèmes d'écriture	Régions	Périodes
Sumérien-Accadien (cunéiforme)	Mésopotamie	3100 avant J.-C. à 75 après J.-C.
Proto-élamite	en Elam (Mésopotamie)	3000 avant J.-C. à 2200 avant J.-C.
égyptien	Egypte	3100 avant J.-C. au II-ème siècle
Proto-indien	au bassin de l'Indus, sous-continent Indien	~2200 à ~1000 avant J.-C.
Crétois (hiéroglyphique, Linéaire A et Linéaire B)	en Crète et en Grèce	2000 avant J.-C. au XII-ème siècle
Hittie et Louwian (hiéroglyphique anatolien)	en Anatolie en Syrie	1500 avant J.-C. à 700 avant J.-C.
Chinois	en Chine	1500-1400 avant J.-C. jusqu'à aujourd'hui

Tab.I.2.1. Genèse de l'écriture

Le premier système élaboré d'écriture est l'écriture cunéiforme (en forme de coin) qui apparut à la fin du IV^{è} millénaire avant J.-C. Cette graphie était employée pour noter la langue des Sumériens qui habitaient la partie inférieure de la Mésopotamie (la terre entre les deux fleuves : le Tigre et l'Euphrate qui s'écoulent vers le golfe Persique).

L'élaboration de la graphie cunéiforme fut le résultat de la nécessité économique et les tout premiers documents élamites et sumériens ne concernent pas la communication au sens habituel du terme (ou mythe oral ou une composition poétique), mais c'étaient de simples listes d'objets notés de façon figurative avec le nombre de chacun des objets associés en général à d'anciens centres de culte ou de cour. Ils sont des registres purement économiques ou administratifs.

I.2.2. Le développement de la transcription phonétique

En théorie, les signes correspondent à des mots isolés qui peuvent fournir un équivalent relativement exact du signe et du son, de l'image et de la parole.
Selon ce qui a été observé dans les graphies de l'Asie occidentale, la cunéiforme de Mésopotamie, les hiéroglyphes égyptiens et l'anatilien, le développement de l'indicateur sémantique n'était pas prononcé, mais avait servi de distinguer les signes indiquant des mots qui avaient plus d'un sens (ambiguïté sémantique).

Par exemple, en cunéiforme le signe pour le mot "AŠŠUR" veut dire la "cité" et le "dieu patron". Un déterminatif peut être ajouté au signe initial pour indiquer à quelle classe il appartient et ce sera soit le signe pour "cité" soit pour "divinité".

L'emploi de tels déterminatifs pour limiter l'ambiguïté des signes ajouta à l'écriture une nouvelle complexité bien que cet emploi ait aidé l'interprétation des signes existants. Toutefois, l'évolution la plus importante qui ouvrit la voie à l'alphabet moderne consiste à l'écriture syllabique et l'emploi systématique du principe phonétique. Les signes n'ont plus besoin de

distinguer entre les sens séparés, mais ils peuvent noter le son lui-même pour tenir compte du sens.

Par exemple, le mot sumérien "TI" (= la vie), qui est difficile à mettre sous forme figurative, peut être exprimé au moyen du signe désigant "flèche" qui est aussi "TI".

Ce changement implique l'emploi préférenciel de l'indicateur d'équivalence phonétique sur celui de la sémantique. Cette nouvelle méthode de représentation est plus abstraite pour la transcription et permet d'économiser l'écriture. Les indicateurs phonétiques furent employés avec des signes transcrivant les mots (comme l'étaient les signes sémantiques) pour spécifier la façon dont le signe devait être prononcé. De tels mots sont découpés en leurs syllabes à l'aide de signes rendant ces mots déjà présents dans la langue : combinaisons variées de consonnes et de voyelles, d'arrêts et de respiration.

L'emploi et la combinaison de signes utilisant des mots et des syllabes donnèrent naissance à des syllabaires. Ces derniers basés sur le principe phonétique employaient un nombre de signes très réduit. Cette évolution eut lieu aux franges des grandes civilisations :

- Le japonais élabora un syllabaire employant des signes chinois qui comprenaient certains signes phonétiques ;
- Les Elamites et les Hourrites firent de même avec le sumérien ;
- L'égyptien peut-être vu comme apparenté aux syllabaires ouest-sémitiques qui sont les ancêtres de l'alphabet.

Les systèmes d'écriture syllabique emploient un jeu limité de signes et sont relativement faciles à apprendre et à utiliser. Un avantage considérable pour les missionnaires, pour les explorateurs du monde ou pour les commerçants depuis l'antiquité qui préparaient des systèmes d'écriture pour communiquer avec des peuples externes.

I.2.3. L'alphabet

Il y a deux avis sur l'invention de l'alphabet : - le premier veut qu'il ait été inventé en Grèce vers 750 avant J.-C., dans la période qui précéda immédiatement les grandes réussites ioniennes et athéniennes; - le second, qu'il ait été inventé par les Sémites occidentaux vers 1500 avant J.-C.

Selon J. Goody, il considérait que les deux opinions sont correctes, car la première s'appliquait à un alphabet complet avec consonnes et voyelles, la seconde à l'alphabet comportant les seules consonnes.

Qu'est-ce que l'écriture alphabétique a facilté?
L'écriture à ses débuts fut mise au service de l'économie, la politique, et la formation des scribes le tout en relation avec les temples. La complexité des systèmes logographiques et le désir de contrôler l'instruction voulait dire que la connaissance de l'écriture était réservée à une fraction ou une classe de la population. L'une des tâches du cunéiforme fut la notation d'informations sur les mouvements des corps célestes qui servit de base aux progrès ultérieurs en astronomie et les mathématiques. Un tel processus n'imposait pas l'emploi d'un système de nature linguistique, car les mathématiques demeurent un système logographique plus qu'alphabétique.

L'invention de l'alphabet incorporant le système syllabaire, amena une réduction considérable du nombre de signes et l'emploi d'un système d'écriture potentiellement illimité à la fois dans sa capacité de transcrire la parole et dans son accéssibilité.
Les dérivés de l'alphabet syllabaire cananéen gagnèrent rapidement l'Europe et l'Asie et atteignirent plus tard les autres continents, rendant accessible une graphie facile à apprendre et facile à employer.

I.2.4. Ecriture et progrès

Nous avons traité l'alphabet comme une invention exceptionnelle, tout en employant de bonnes raisons historiques et empiriques de ce qu'on connaît des historiens de l'alphabet. Mais la raison essentielle porte sur des raisons théoriques, car l'alphabet est l'oeuvre d'une création abstraite. Par exemple, dans le cas de l'anglais, une subdivison de 21 élements consonantiques et de 5 voisés.

L'alphabet représente aussi le point d'aboutissement d'un processus d'élaboration "logique". Selon J. Goody, on arrive au niveau d'abstraction ([Goody, 94], p.75) à partir :

1. de logogrammes (des signes pour des mots),
2. de syllabes,
3. de l'initiale, une lettre ou une nouvelle lettre pour représenter l'élément phonétique commun à un groupe de syllabes.

L'alphabet ne peut émerger que lorsque toute la série des syllabes est mise dans une forme graphique.

Le système syllabaire en tant que système d'écriture a de nombreux avantages pour le processus d'apprentissage de l'écriture et de la lecture. Une telle occurrence dans la société semble le moyen de stimulii avec des sociétés exterieures possédant ou employant déjà l'écriture.

Toutefois, l'emploi de ce nouvel système oblige à sauter vers un niveau d'analyse abstraite impliquant des opérations antérieures liées au système et systématisant les formes et les origines du savoir.

L'invention de l'écriture amena un admirable instrument d'engagement, de précision et d'analyse conceptuelle, fournissant des changements révolutionnaires dans la culture et l'émergence d'une classe spécialisée dans la technique et l'art de faire des approches intellectuelles sur la réalité.

I.3. Images : Proto-écriture ou écriture en images ?

L'emploi des représentations figuratives sous formes de séquences, comme celles qu'on retrouve en Amérique du Nord, a reçu le nom *d'écriture en images*. Un des premiers experts en matière de pictographie, Garrick Mallery (1886), a décrit cette écriture comme un mode de transmission, d'enregistrement d'idées et d'évènements par des moyens graphiques. C'est une forme d'écriture de la pensée qui s'adresse directement à la vue. Comparativement, une autre forme similaire d'une telle communication étant le langage gestuel (communication par des gestes).

Toutefois, une distinction entre l'image isolée et l'emploi de séquences dans le dessin graphique s'impose. Un exemple de pictogramme equivalent à un signe se suffit à lui-même à la fois

physiquement et morphologiquement, car il fait partie d'un système sémiotique restreint, comparable à d'autres signes semblables et interprétables.

Un système séquentiel de pictographes tel que retrouvé dans l'écriture "embryonnaire" est très proche de *la bande déssinée*. Par exemple, les représentations picturales du mythe national de la fondation du royaume de l'Ethiopie, qui inclut la fameuse visite de la reine de Sabâ et Salomon, visite qui eut pour résultat la naissance de Ménélik, le premier souverain de la lignée de Judah.

Deux opinions sur l'écriture pictographique s'opposent. D'une part, G. Mallery et beaucoup d'autres pensent que la pictographie est une *écriture de la pensée* (système idéographique) qui représente des objets, des évènements, des concepts ou des pensées, sans intervention du langage et par conséquent des éléments linguistiques. Cette représentation est aussi maintenue dans le langage par gestes.
Et d'autre part, J. Goody et al. parlent de séquence graphique comme une suite d'éléments distincts, qui peuvent naturellement être figuratifs. Dans ce cas, l'écriture pictographique implique un rapport systématique reliant des motifs les uns aux autres. Dans ces conditions, les systèmes de codage et de décodage des humains doivent inclure une composante linguistique comme dans tous les procédés de pensée ou de conceptualisation. Et selon cette thèse, une définition importante des gestes symboliques, les voit comme des actes qui ont une traduction verbale directe.
Les opinions sur la seconde thèse, pour mesurer la profondeur du langage enfoui dans l'activité spécifique humaine où le langage apparaît comme un intermédiaire essentiel, qualifient l'art figuratif de *l'inséparable du langage* et que le développement des formes graphiques soit lié à la parole.

> Selon J. Goody : "*Ni dans l'imagination ni sur la toile, une représentation du monde extèrieur ne peut être indépendante de l'usage linguistique, c'est-à-dire indépendante non seulement des catégories mais de l'expérience accumulée et incorporée dans le discours*." [Goody, 94] (p.27).

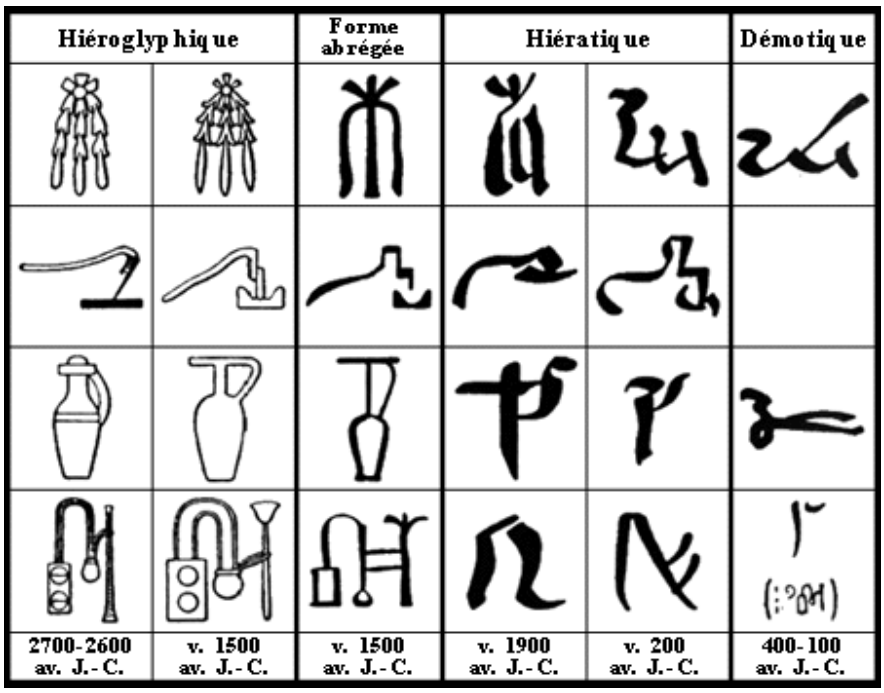
Le dessin et le langage sont souvent vus comme des modes concurrents de communication [Lieberman, 72]. Tout usage élaboré de la représentation visuelle exige un système conceptuel développé (humain) qui est essentiel à l'emploi du langage. Dans le langage, le mot et l'image sont complémentaires. Les aspects "communicatifs" ou "expressifs" dans les premiers stades de *l'Homo sapien* pourraient être décrits comme un embryon d'écriture. Un tel système d'écriture ne conduit pas à une quelconque sémiotique formelle, mais à l'implication d'un système élevé d'élaboration conceptuelle qui soit doté d'une sémiotique simple et générale.

I.4. Illustrations : systèmes d'écriture en image

I.4.1. LES HIEROGLYPHES

Le mot «hiéroglyphe» signifie littéralement «sculpture sacrée». Historiquement, les Égyptiens ont utilisé les hiéroglyphes exclusivement pour des inscriptions sculptées ou peintes sur les murs des temples. On retrouve aussi cette écriture picturale sur des tombes, des papyrus, des planches de bois recouvertes de stuc, des tessons de poteries et des fragments de calcaire.
Les hiéroglyphes (*Tab.I.4.1.*) sont une forme originale d'écriture dont découlent deux formes nouvelles d'écriture qui furent qualifiées de hiératique et de démotique. L'écriture hiératique était

une forme simplifiée des hiéroglyphes utilisée à des fins administratives et d'affaires, et pour des textes littéraires, scientifiques et religieux. Le mot «démotique», qui en grec signifie «populaire», désigne l'écriture employée pour les besoins de la vie de tous les jours. Au cours du III^{e} siècle après J.-C., les hiéroglyphes furent graduellement supplantés par l'écriture copte.



Tab.I.4.1.
Des hiéroglyphes et ses équivalents dans l'écriture cursive

Prise de : G. Steindorff and K. Seele, *When Egypt Ruled the East*, Chicago : 1942, p.122

L'écriture hiéroglyphique, en tant qu'image, est soumise lorsqu'elle véhicule du langage à une double contrainte [EGYPTE, 98] :

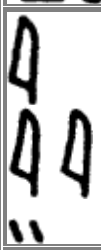
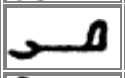
- le calibrage, qui impose au hiéroglyphe du scarabée la même taille qu'à celui du vautour ou de la pyramide,
- l'orientation : les signes représentant humains et animaux ont le regard tourné vers le début de l'inscription, indiquant le sens de lecture.

Les Égyptiens ne se sont pas souciés de se servir de la réduction alphabétique car pour eux l'écriture n'est pas une simple technique permettant de noter la langue, elle est avant tout une image du monde, un art du visible qui assure à ce qu'elle peint l'immortalité.

I.4.1.1. Alphabet hiéroglyphique

Parmi les phonogrammes, 24 signes-consonnes constituent ce qui aurait pu devenir un «alphabet».

Tandis que les caractères cunéiformes évoluent vers des formes anguleuses abstraites [CUNEIFORME, 2001], les hiéroglyphes conservent au long de leur histoire toute leur beauté figurative (*Fig. I.4.1.*). Ils ont en outre une efficacité magique : les caractères désignant le nom d'une personne étaient censés contenir son identité [ROBINSON, 95]. L'écriture avait le double pouvoir d'évoquer réellement et de faire passer à l'immortalité [EGYPTE, 99].

Signe	Objet représenté	Son approximatif
	Vautour percnoptère	*aleph* ([*a*] sémitique)
	roseau fleuri double roseau fleuri double trait oblique	*yod* ([*ya*] sémitique) *double yod*
	avant-bras	*' ayin* ([*'ayïn*]sémitique)
	petite caille abréviation hiératique	*ou* ([*W*] sémitique)
	pied	*b* [b]
	siège	*p*
	vipère à cornes	*f*
	chouette côte de gazelle (?)	m
	filet d'eau couronne rouge	*n*
	bouche	*r*
	cour de maison	*h (aspiré) [ha]*
	écheveau de lin tressé	*h (*emphatique)
	placenta	*kh* guttural *[kh]*
	ventre et queue d'un mammifère	*ch* [*ch*]
	verrou étoffe pliée	*s*
	bassin d'eau	*ch*
	dune	*q*

	corbeille à anse	*k*
	support de jarre	*g*
	galette de pain pilon	*t*
	entrave (corde)	*tch*
	main	*d*
	serpent	*dj*

Fig. I.4.1. D'après l'oeuvre de Christiane Ziegler, « Naissance de l'écriture », RMN, 1982.

I.4.1.2. Les signes : logogramme, phonogramme et déterminatif

Trois types de signes, dont les valeurs se complètent et souvent se redoublent, coexistent dans l'écriture égyptienne :

Logogrammes	**Phonogrammes**	**Déterminatifs**
un signe = un mot	un signe = un son procédé du rébus	précisant dans quelle catégorie d'objets ou de concepts le signe doit être classé.
Exemples :		
 signifie "soleil" et par extension "clarté ", "jour", " moment ", etc.	 signe de la "bouche", se prononce "er", sert à noter la consonne "r".	 indique l'idée du mouvement.

Tandis que les caractères cunéiformes évoluent vers des formes anguleuses abstraites, les hiéroglyphes conservent au long de leur histoire toute leur beauté figurative.

I.4.2. LE SUMÉRIEN

Bien qu'il y ait des points de désaccord concernant l'origine des séquences type sumériennes, une zone sur laquelle il n'est aucun désaccord est que la séquence type "n'a pas été inventée" dans le quatrième millénium avant J-C. Les séquences type sumériennes les plus anciennes ont été découvertes en Uruk en 1924 par un groupe d'archéologues allemands menés par Jules Jordan.

Ces textes ont été trouvés à la strate d'Uruk IV (*Fig.I.4.2a.*) et ont été donc datées 4100 et 3800 avant J-C. Approximativement 1000 textes contenant beaucoup de pictogrammes concrets.

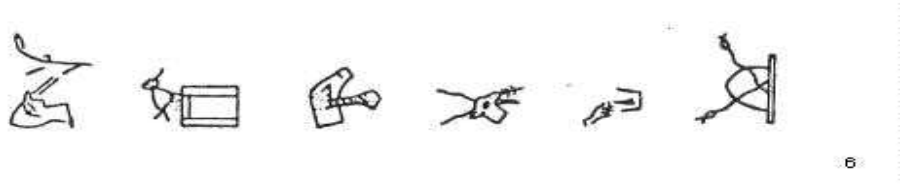


Fig.I.4.2a. D'après l'oeuvre de I. J. Gelb, A Study of Writing, Chicago: Univ. of Chicago Press, 1956 p. 66.

Dans des tablettes d'argile, un certain nombre de pictogrammes (pictographes) se rapportent à des objets tels que du bétail. I.J. Gelb a formulé une hypothèse au sujet de leur origine : dans son étude (« A Study of Writing », 1956), la transmission écrite évolue toujours du simple au complexe.

Le dessin (ou graphie) est d'abord réalisé pour l'esthétique ou pour des buts religieux. Ensuite, il est adapté à identifier les dispositifs mnémoniques. En exemple, les marques de maçons pour identifier différents types de coupes employées par les pierre-coupeurs. Ces marques peuvent être lues par les disciples, mais n'ont pas une grammaire ou un vocabulaire.

I.4.2.1. Ecriture en image à sa forme abstraite

Le symbole a changé les dispositifs mnémoniques en mot, pour devenir plus abstrait dans l'aspect. L'évolution d'une variété de symboles est décrite dans le diagramme ci-dessous :

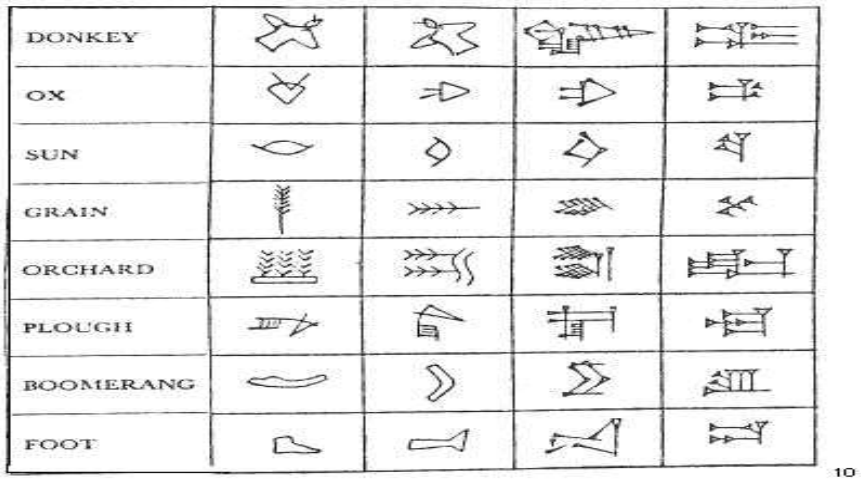


Fig.I.4.2b. D'après l'oeuvre de I. J. Gelb, A Study of Writing, Univ. of Chicago Press, 1956, p. 37.

Dans la strate d'Uruk II, Gelb montre des logographes (écriture en image, *Fig.I.4.2b.*), et idéographes (symboles pour les objets concrets), tels qu'une croix cunéiforme pour l'étoile signifiant également dans sa forme abstraite le «dieu du ciel».

L'étape syllabique, selon Gelb, est indiquée par des dispositifs phonétiques (le mot "homme" symbolisé par une couronne est utilisé en tant qu'élément du nom " Su-en " ou " péché "). Ces dispositifs sont trouvés dans les niveaux de strates d'Uruk IV et III. Il a noté que l'évolution de l'inscription du pictographe à la syllabe a été produite très rapidement.

Le **premier alphabet organisé connu** est en écriture cunéiforme simplifiée de trente signes. Il fut inventé à **Ougarit (***Fig.I.4.2c.***)**, ville commerçante de la côte syrienne vers le **XIVème siècle avant J.-C.** et servit à noter la langue sémitique locale. C'est dans cette écriture cunéiforme alphabétique que les habitants d'**Ougarit** ont écrit leurs mythes, leurs rituels religieux, une partie de leur correspondance et les textes administratifs du royaume (*Fig.I.4.2d.*).

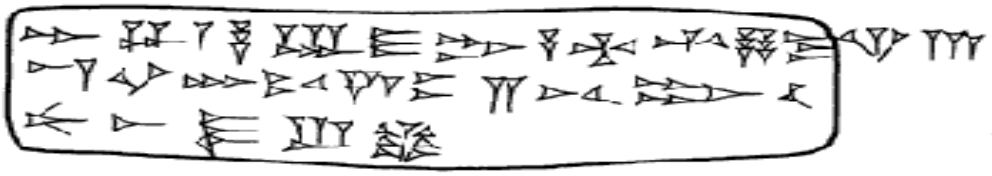


Fig.I.4.2c. Abécédaire d'Ougarit (D'après Virolleaud, Syria, XXVIII, p. 22)

I.4.2.2. Alphabet Sumérien selon le codex de Hamurabi

Signe	Objet représenté	Son approximatif
	NA	*na*
	A	*a*
	Š A(Sha)	š*a (sha)*
	Š U	š*u (shu)*
	GAL	GAL, (*Rabû=*`great')
	KI	*ki, ke, qí*
	MU	*mu*
	MA	*ma*
	BI	*bi*
	AN, DINGIR	*an*
	KÁM	
	I	*i* = `five'
	IM	*im*
	Ú	
	Š I	š*i*, š*e*, *igi*, *lim*
	BAD	*be, bad/t/T*
	RI	*ri, re*
	IR	*ir, er*
	RA	*ra*
	UD, U	U = UD = *ummu* `day'

	DIM	*dim, tim, Tim*
	NI	*ni, né, lí, lé, ì, zal*

Fig.I.4.2d. Signes triés en diminuant la fréquence de l'occurrence dans le codex Hamurabi (1850 avant J-C.)

I.4.3. LE CHINOIS

Compte parmi les écritures en usage d'aujourd'hui, l'écriture chinoise est la seule qui ait traversé autant de millénaires.
Il s'agit déjà d'un système pictographique cohérent, bien que les graphies ne soient pas encore stabilisées (Fig.I.4.3.).

Fig.I.4.3.
Inscriptions oraculaires, Chine, XIIème s. avt J.-C.
Paris, BnF, manuscrits orientaux

Au IIIe siècle avant J.-C., l'empereur Qin Shihuang, s'intéressant à l'écriture dans sa politique d'unification de la Chine, demande à son ministre Li Si de mettre un terme à la prolifération anarchique des caractères. Li Si établit une liste de 3 000 caractères dont il fixe la forme.

Cependant, le nombre de caractères ne va cesser d'augmenter :

- 8 000 à la fin du Ier siècle de notre ère,
- 18 000 au IIIe siècle,
- 30 000 au XIe,
- 47 000 au XVIIIe,
- environ 55 000 aujourd'hui dont 3 000 sont d'usage courant.

Cette tendance à la prolifération, répondant à un enrichissement quasi permanent du lexique, est rendue possible par la nature même des caractères. C'est ce qui explique la grande pérennité de cette écriture.

Les caractères chinois peuvent jouer deux rôles : celui de composant sémantique (radical) et celui d'élément phonétique qui sert à indiquer la prononciation du caractère. Il existe 227 clés ou radicaux dont la présence indique la catégorie de choses, d'idées, etc.

On distingue traditionnellement huit traits (*Tab. I.4.3.*) de base en chinois :

Trait	Son	Signification

	diǎn	le point
	héng	le trait horizontal
	shù	le trait vertical simple
	piě	le jeté
	duǎnpiě	le jeté court
	nà	l'appuyé
	tí	le relevé
	shùgōu	le trait vertical en crochet

Tab. I.4.3. Traits de base de l'écriture chinoise.

I.4.3.1. La genèse des caractères chinois

Les caractères chinois s'analysent en –*Figures simples*, qui peuvent être des images, ou des symboles, et en –*Figures composées,* qui peuvent être des agrégats logiques, ou des complexes phoniques.

Figures simples : images

Les images sont une représentation stylisée des objets, sans la moindre indication de la prononciation.

Poisson	Arbre	Remarques
		le dessin initial transcrit le poisson / l'arbre en figurant, (inscription oraculaire)
		le tracé évolue peu

		la forme manuscrite ancienne est définitive,
		forme moderne.

- **Figures composées : agrégats logiques**

La grande majorité des caractères chinois ont été constitués par l'association de plusieurs caractères : ils sont formés par l'agglomération de plusieurs caractères dont les sens se conjuguent.

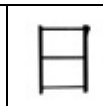
Exemple	Figure
le pictogramme « *soleil* »	
redoublé, produit *chang*, signifiant « brillant », « glorieux »	
triplé en pyramide, il génère *jing*, signifiant « l'éclat de la lumière »	

I.4.3.2. Les complexes phoniques

Aucun élément du caractère ne donne d'indication sur sa lecture. Les complexes phoniques sont de loin les plus nombreux. Ils associent deux éléments graphiques dont :

— l'un indique le sens ou plutôt la famille lexicale à laquelle appartient le caractère,
— l'autre la prononciation.

Exemple sur le sens			
le pictogramme du « *soleil* »			
pourra être associé au caractère *wang*, « le roi »		pour noter le mot wang « brillant », « éclatant »	
Exemple sur la prononciation			

ce même pictogramme « *soleil* »			
flanqué du caractère *gan*, pictogramme figurant un pilon		utilisé ici pour indiquer la lecture du nouveau caractère ainsi formé, *gan* « le soir », « le soleil couchant »	

Le pictogramme de l'*arbre*  intervient :

- dans les noms d'arbres :


song, le pin


bo, le cyprès


zong, le palmier

- dans les noms des objets en bois :


ban, la planche


pa, le râteau


zhuo, la table

I.5- Les mots et les figures : apport de la rhétorique

Les mots visualisables sont ceux dont les signifiés sont perceptibles par le sens de la vue. Des études sur les images mentales nous confirment, qu'à l'intérieur du registre verbal, la fréquence des noms concrets est supérieure à celle des noms abstraits dans tout type d'apprentissage associatif, sériel, discriminatif, etc., et cela sur le même type de tâche [DENIS, 79]. Ce qui confirme l'ouverture du développement de plusieurs méthodologies dans la construction de systèmes d'information iconographiques et sur le multimédia en général [PENTLAND, 94], [TURNER, 96], [PICARD, 96], [THIVOLLE, 98].

Des travaux remontant à l'antiquité gréco-romaine, mais aussi des travaux scientifiques contemporains, confirment cette hiérarchie de souvenirs visuels et abstraits. Cette supériorité de la mémoire visuelle est un constat de grande importance : quand l'information initiale est de nature visuelle, la mémoire de l'observateur retiendra le dessin, les couleurs, les traits… aisément [RADA, 95], [MUSGRAVE, 96], [NASTAR, 97]. Mais aussi indirectement, si l'information initiale est de nature linguistique, le lecteur remémorerait facilement les mots correspondant à des objets concrets qu'à ceux sans correspondance visuelle.

I.5.1- Démarche des rhétoriqueurs greco-romains

Dans une démarche associative à relier un concept purement abstrait à un objet concret visualisable : une image, les rhétoriqueurs grecs et romains doués d'une mémoire « prodigieuse » : non pas innée mais travaillée à partir de principes, qui restent valables, sur la force des figures de style : la *comparaison*, l'*exemple* et de ces deux figures précisément de la rhétorique : la *métonymie* et la *métaphore*.

La métonymie consiste à employer un nom pour un autre qui lui est lié : un *signifié* étant désigné par un *signifiant* qui ne le désigne pas d'ordinaire. Cette attitude renforce la visualisation d'une expression en remplaçant le signifiant habituel *abstrait* par un signifiant voisin mais plus *visualisable* et *concret*.

La métaphore est une comparaison ou un transfert de sens en substituant une expression abstraite par une expression concrète (ou inversement). Dans les termes de la substitution, les liens ne sont pas évidents : la puissance du langage concret dans un procédé de visualisation peut marquer une stylistique littéraire, scientifique ou technique.

La genèse du message oral ou écrit relève des commandes du *rhéteur* pour ses discours ou plaidoiries. Rappelons que dans la rhétorique gréco-romaine, le message comprenait cinq parties :

- *l'invention :* recherche et découverte des matériaux qui seront utilisés dans le discours,
- *la disposition :* recherche du plan,
- *l'élocution :* rédaction soit mentale, soit écrite du discours,
- *la mémoire :* mise en pratique des techniques de l'art de la mémoire pour retenir le texte afin de prononcer le discours ou la plaidoirie sans notes,
- *l'action :* les voix et le geste pendant le cours ou la plaidoirie.

A l'issue de cette discipline, la rhétorique et des traces retrouvées dans les textes anciens sur l'art du discours et de présentation d'une plaidoirie comme chez les rhéteurs *Quintilien* et *Cicéron*.

Nous présentons quelques modèles (*Tab.I.5.1a*) issus des textes anciens :

Genèse du discours rhétorique (Modèle plan)			
étapes	Modèle d'*Herennius* (*86 avant J.-C.*)	Modèle de *Cicéron* (50 *avant J.-C.*)	Modèle des *Jésuites* (*quelques siècles plus tard*)
1	Proposition	Exorde	Eloge
			Paraphrase
2	Preuve	Narration	La cause
			Le contraire
			Le semblable
3	Confirmation de la preuve	Confirmation	L'exemple
4	Mise en valeur	Réfutation	Témoignages des anciens
5	Résumé	Péroraison	Epilogue

Tab.I.5.1a. Genèse de quelques modèles.

Rien de périmé dans ces modèles et des étapes de classifications dans la fabrication de discours. On peut bien s'objecter qu'il s'agit là du cas des expressions orales : discours ou plaidoiries, cas très particulier par rapport au contexte du texte écrit. Mais, il faut savoir que les rhéteurs écrivaient préalablement leurs prestations, tel Quintilien ou Cicéron. Et d'autre part, à l'oral ou à l'écrit, il semble que les mêmes mécanismes de persuasion ou de narration soient valables.

Plus tard, d'autres modèles sont nés pour s'adapter au contexte de nouvelles applications et dans des situations nouvelles de la communication. Je citerai ainsi à titre d'exemple trois modèles plans contemporains (*Tab.I.5.1b.*):

Les grandes lignes de ces modèles étant définies, il reste à son utilisateur de remplir ceux-ci avec les titres, les mots et les informations se rapportant à son projet (pas de règles contraignantes, mais en fonction des circonstances du projet, de son thème, et de l'importance quantitative et qualitative du sujet).

Modèles plans contemporains			
étapes	Modèle *compte rendu scientifique*	Modèle *SPRI* *(ou de Louis Timbal-Duclaux)*	Modèle de *Laswell*
1	Problème	Entrée en matière	qui
	Travaux antérieurs		
2	Protocole expérimental	Situation	a dit quoi
		Problème	à qui
3	Résultats	Résolution	pourquoi
4	Interprétation et discussion	Information	où
			quand
5	Conclusion et recherches futures	Terminaison	avec quel résultat

Tab.I.5.1b. Quelques modèles contemporains.

I.5.2- Ecrire le plan

Cette deuxième partie est consacrée à l'écriture et à la réalisation d'un modèle plan dans le contexte de production d'un message : il s'agit de la transposition de l'écriture dans une grille d'analyse selon un modèle plan. Cette activité [RICHAUDEAU, 92] peut s'analyser comme un processus suivant cinq thèmes (*Tab.I.5.2.*) :

Processus d'activité de l'écriture		
étape	*thème*	*contenu*
1	Projet	Celui d'imaginer ou de développer une idée, un thème ou une description. A ce stade rien n'est encore rédigé : *le projet mental*.
2	But	Transmettre ce projet à des lecteurs qui seront des usagers potentiels du produit de ce projet : *la cible*.
3	Moyen	Pour transmettre ce projet, le langage est la concrétisation par des signes matériels des idées découlant de ce projet mental et destinées à la cible. Cette représentation complexe relève de *4 opérations* : – le codage lexicologique, – le codage syntaxique, – le codage rédactionnel et – le codage typographique.
4	Moindre effort	Conduit aux choix des mots et leurs assemblages en phrases afin que le lecteur (la cible) parcoure le texte construit avec le minimum d'effort de compréhension : *les lois de la lisibilité*.

5	Exploration	Suivant l'objectif de l'auteur du projet envers ses lecteurs (la cible), le moyen d'évoquer des sentiments, de transmettre des informations, de décrire un événement, un objet, etc. avec le moindre effort : *le style* qui s'ajoute à une écriture fonctionnelle.

Tab.I.5.2. Plan Richaudeau.

Cette grille est valable pour l'acte d'écriture selon un modèle plan dans un processus communicationnel général, qui fait intervenir des facteurs (*Fig.I.5.2.*) : un projet mental, une cible de lecteurs, un codage linguistique, une recherche de lisibilité et une recherche de style.

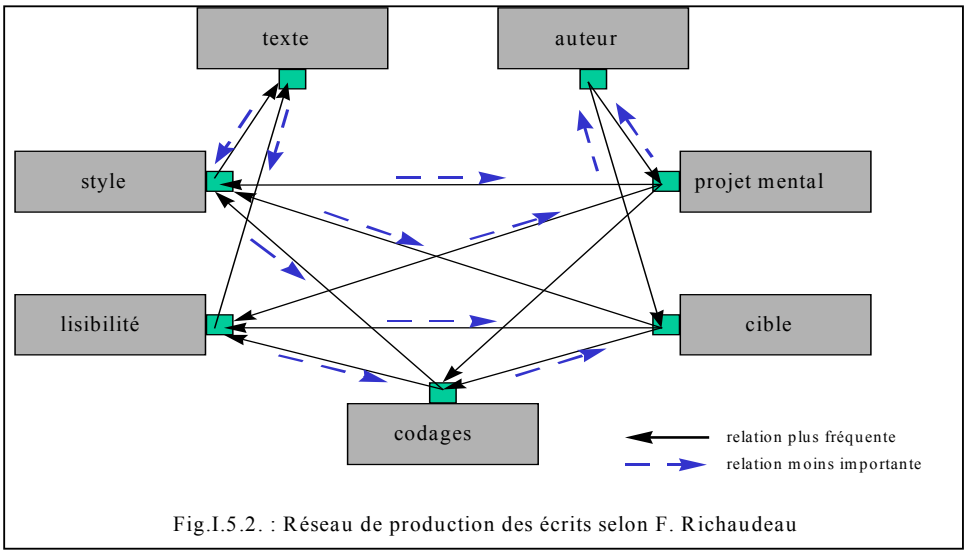


Fig.I.5.2. : Réseau de production des écrits selon F. Richaudeau

Ces facteurs se combinant entre eux par des relations non-linéaires d'actions directes et de rétro-actions, et peuvent être représentés par une structure en réseau : une considération prise en permanence à tous les niveaux d'étude et de pratique du processus d'écriture.

I.6. Conclusion

Dans le contexte de l'analyse des contenus multimédia et particulièrement la recherche d'un lien commun entre le graphique (l'image), le verbe (la parole) et le texte (l'écrit), il m'a paru évident de remonter dans le temps et de rechercher les origines de l'activité langagière qui est source des connaissances et du savoir humain.

A travers notre étude basée sur les travaux de l'anthropologue Jack Goody, la base physique de l'écriture reste clairement la même que celle du dessin, de la gravure et de la peinture : les arts

graphiques. Si l'écriture a ses racines dans les arts graphiques, dans des dessins signifiants (des indices figuratifs et d'autres arbitraires de type abstrait), le niveau d'abstraction de l'écriture en représentant le monde est loin de rester graphique.

La graphie donne accès d'une façon très différente à la lecture et à la composition d'oeuvres écrites. Cette notion de maîtrise de l'écriture comme indicateur de savoir entraîne le même système d'évaluation dans le cas d'une écriture syllabique ou alphabétique.

Un objet dessiné peut signifier une sorte d'objet en description graphique et en même temps qu'il indique pour cet objet un morphème ou une description verbale plus étendue. Or, le langage est arbitraire, cette hypothèse reste valide pour l'écriture graphique, c'est-à-dire qu'un lien pictographique direct soit possible dans le cas de certains signes (logographiques) pour faire la correspondance à des mots ou à des morphèmes. Cette correspondance fait le lien aux cultures orales et aux cultures écrites.

Le progrès principal, qui a été observé par les scientifiques en la matière, réside dans le niveau auquel le système graphique réussit à doubler le système linguistique : – la correspondance sémantique entre le mot et le signe, puis – la correspondance phonétique. Une telle présentation des signes figuratifs était incorporée dans tous les premiers systèmes d'écriture. Mais ils n'étaient pas la seule source des signes pour établir la correspondance entre la langue et la graphie : appelée écriture.

Dans les systèmes d'écriture syllabique ou alphabétique, la référence première d'un signe est liée au son, alors que dans les écritures ou formules figuratives la référence la plus immédiate est liée à l'objet ou à l'événement lui-même.

En résumé, dans cette étude, nous avons cherché à écarter un point de vue émis sur l'analyse de l'image que « Jamais aucune explication d'image ne peut rendre compte de tout ce que contient un texte » ; Le seul équivalent de l'image demeure l'image elle-même [BAXTER, 96].

Quant aux différents types d'écriture, il convient de suspendre son jugement. Ce que l'on a gagné en abstraction et simplicité avec l'alphabet, ne l'a-t-on pas perdu en qualité iconique et graphique des images ? Aujourd'hui en laissant de côté les civilisations antiques, que dire de l'écriture chinoise qui continue toujours d'employer ses logogrammes (sans transformation de type alphabétique) ?

Nous avons cherché à travers cette étude, qui puise ses sources dans des recherches anthropologiques, culturelles, communicationnelles, langagières, et l'apport de la rhétorique dans la représentation du discours, de restituer un aspect fondamental sur l'analyse de l'image [BOUSQUET, 91] et du multimédia en général. Dans ce processus d'analyse, il existe une association image-écrit et une production de l'écrit à partir de l'image.

Ainsi, nous convenons à émettre un avis sur les différents canaux communicationnels comme sources de connaissances et l'établissement d'un lien commun entre eux :

- les arts graphiques donnent accès à la lecture et à l'écriture,
- l'écriture graphique ou imagée est la source des écritures syllabiques ou alphabétiques,

- l'écriture a ses racines dans les arts graphiques,
- le système graphique réussit à doubler et à amplifier le système linguistique,
- les sources de connaissances visuelles ne sont pas indépendantes ou infranchissables des sources langagières, en exemple : l'apport de la rhétorique.

L'analyse du multimédia subordonnée à l'analyse de son contenu textuel constitue une nouvelle source de connaissance complémentaire à la nature du document source. La description textuelle selon un processus adéquat de production des écrits constitue une autre forme d'accès à une source de connaissances qui n'est pas forcément écrite.
La nouveauté dans cette prise de conscience sur le multimédia et ses extensions dans l'écriture [SCHMUCK, 95], [ROMER, 95], tranche en faveur de l'exploitation générique de l'écrit. La construction des résumés est désormais un enjeu non négligeable exprimant de véritables connaissances documentaires [ROUSSEAU, 94], [POULAIN, 96], [PRIÉ, 2000]. La relation entre document et son résumé textuel dans les réseaux documentaires donne accès à une double perspective d'analyse soit sur le document (document primaire) comme tel, soit sur la sémantique explicite de son contenu (document secondaire : résumés, annotations, descriptions, thèmes, titres, etc.).

La suite du travail propose l'écrit comme source thématique offrant les concepts opératoires et organisationnels sur l'indexation automatique des documents multimédia.

CHAPITRE II :

État de l'art sur l'indexation documentaire

II.1- Introduction

Les bases de données documentaires ont l'ambition de mémoriser des informations sur les contenus des documents en fonction de plusieurs critères et dimensions. Afin de répondre aux interrogations des usagers, cette base leur fournit une sélection de documents.

En général, l'usager de la base ne connaît pas de références susceptibles de l'intéresser. Il essaye de formaliser sa demande lors de l'interrogation par le travail sur un thème. Il constitue par la recherche un dossier d'information rassemblant l'ensemble des documents s'y rapportant.

Ainsi, les systèmes d'information documentaires ont pour but de répondre à une telle demande d'information en fournissant les documents adéquats que le système retrouve grâce à une ***indexation «*** judicieuse ***»***.

L'opération d'indexation est particulièrement difficile dans la mesure où elle pose le problème de la représentation du sens du texte. Dans ce cas précis, il faut souligner que les linguistes, les statisticiens et les chercheurs en informatique documentaire la traitent différemment.

II.2- Fondements théoriques des langages d'indexation

Cette partie constitue une synthèse des recherches consacrées aux fondements théoriques des langages d'indexation.

Le terme *langage d'indexation* compte parmi ses synonymes :

- *langage documentaire*, *langage contrôlé*, etc.
- et ce terme recouvre également de nombreux équivalents anglais : « *indexing languags* », « *documentary language* », « *information retrieval language* », etc.

Le langage d'indexation est un langage artificiel, c'est-à-dire construit à l'aide d'un ensemble de règles données, servant à la représentation abrégée du contenu d'un document [RIVIER, 90]. Dès lors, l'indexation consiste à détecter les termes les plus représentatifs du contenu du document, ce qui peut être considéré comme une forme d'acquisition de connaissances sur le contenu documentaire [DACHLET, 90], [LAMROUS, 97].

II.2.1- Classification des langages d'indexation

Les langages d'indexation ne sont pas rigoureusement codifiés, mais ils sont répartis en groupes ou classes.
En nous inspirant des travaux de J. Maniez [MANIEZ, 87], nous pouvons les représenter selon deux axes (*Fig. II.2.1.3*) :

- Axe 1 : Langages d'indexation contrôlés
- Axe 2 : Langages d'indexation coordonnés

II.2.1.1- Axe 1 : Les langages d'indexation contrôlés

Les langages d'indexation contrôlés se rapprochent des langages naturels. Un langage d'indexation *peu contrôlé* correspond aux descripteurs choisis librement pour représenter le contenu d'un document.

Les langages d'indexation *plus contrôlés* se différencient nettement de la langue naturelle. Il s'agit des langages d'indexation post-coordonnés et pré-coordonnés.

II.2.1.2- Axe 2 : Les langages d'indexation coordonnés

Ils sont de deux types :

- *Langage d'indexation post-coordonné :*
  La combinaison des descripteurs se fait au moment de la recherche documentaire, au même titre que les thésaurus.
- *Langage d'indexation pré-coordonné :*
  La combinaison des termes est fixée au moment de l'indexation, tout comme les classifications et les langages en chaîne.

II.2.2.- Schéma de Classification

Nous représentons la classification des langages d'indexation selon les travaux de J. Maniez.

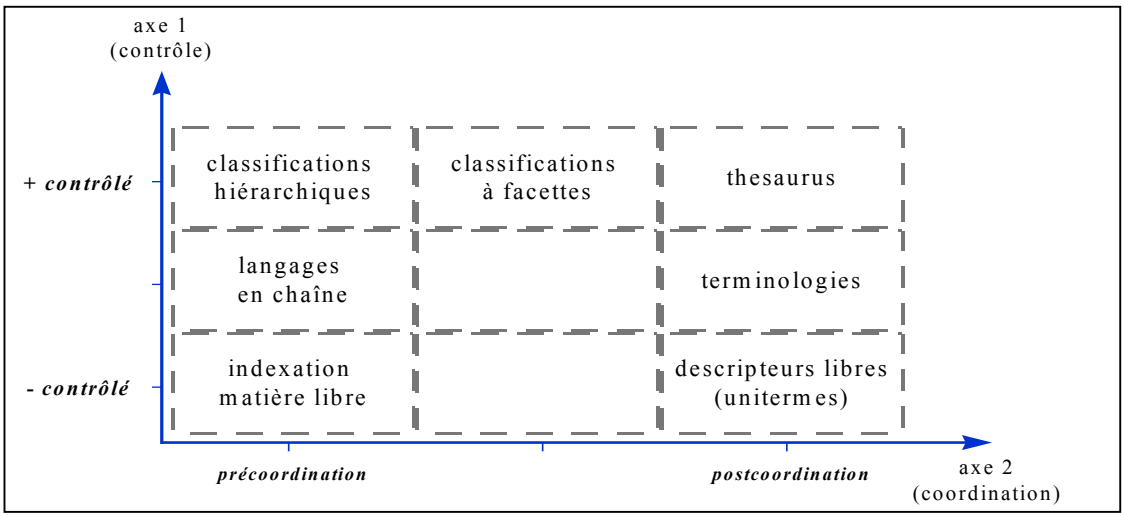


Fig. II.2.1.3. Représentation des langages d'indexation en fonction de leurs degrés de coordination et de contrôle.

II.3- Indexation manuelle

L'indexation manuelle fonctionne comme l'indexation intellectuelle. L'activité de l'indexation d'un document est déléguée à un spécialiste de la documentation.

Le documentaliste décrit les documents pour faire ressortir leur valeur informative. Il accentue particulièrement l'analyse documentaire en utilisant notamment le moyen de la

condensation par l'élaboration de résumés, de synthèses et l'emploi d'un langage de type thésaurus (*Fig.II.3.*).

Soient :
T ou LD : le Thésaurus (T) ou le Langage Documentaire (LD),
M : un mot $\in$ T ou LD,
D : un document à indexer,
alors

index(D) : l'index du document D,
il est défini par : index(D) = {M / M$\in$T ou LD}.

Fig.II.3. L'index d'un document.

Notons, d'une part, que si le thesaurus garantit une certaine normalisation de ses termes, l'indexeur a encore une part de décision importante. D'autre part, le problème d'indexation devient plus complexe lorsqu'il ne s'agit pas d'ouvrages suffisamment généraux pour qu'un classement hiérarchisé par thème soit suffisant.

> En exemple : les revues dont chacune contient plusieurs articles correspondant à de multiples contenus logiques et chaque article traite un problème thématique différent.

Pour mieux résoudre les problèmes liés à l'indexation et à la recherche d'informations, des méthodes ont été développées. Elles consistent à mémoriser les aspects les plus importants du contenu des documents [DEWEZE, 93] et la combinaison logique des descripteurs entre eux.

Ce type de représentation permet :
- d'améliorer l'ensemble des descripteurs caractéristiques du domaine traité pour l'indexation des documents ;
- d'obtenir des réponses plus sensibles aux questions complexes par des combinaisons entre descripteurs.

II.3.1- Indexation par une liste de descripteurs

Ce type d'indexation consiste en la représentation d'un texte intégral par une liste de descripteurs simples (mots simples) ou de descripteurs complexes (ensemble de mots).

Le choix des descripteurs permet de distinguer deux types d'indexations à savoir l'indexation libre et l'indexation par un langage documentaire :

a➢ Indexation libre (I.L.)

Après la lecture du texte, le documentaliste choisit librement des mots ou des expressions qui lui paraissent décrire le contenu du texte.

Les critères de décision sont purement « intuitifs » et « personnels » [LAINÉ, 82]. Aucun calcul précis n'est effectué sur la portée sémantique des mots du lexique et l'emplacement des mêmes mots dans le texte :

- les indexeurs choisiront des descripteurs différents pour le même document et avec peu de descripteurs identiques ;

- le même indexeur traitant le même texte ne choisira pas les mêmes descripteurs à quelques temps d'intervalle selon son humeur et ses préoccupations du moment.

Or, parmi les critères d'indexation à respecter sont :

- la *cohérence* : deux textes ayant trait au même sujet doivent être indexés dans un contexte donné par les mêmes descripteurs [ANDREEWSKY, 77]. Le contexte d'indexation est aussi problématique, car que dire des ***variations temporelles*** du texte ou document ?

- l'*adéquation à la recherche* : utilisation d'un *vocabulaire commun* (termes ou descripteurs) pour l'indexation et l'interrogation.
  Ce *langage commun* et prédéfini (langage documentaire) permet une ***normalisation de l'indexation*** [BASSANO, 81].

b ➢ Indexation par un langage documentaire (L.D.)

Une forme simple du langage documentaire (ou langage contrôlé) consiste en une liste nominative et limitée de descripteurs, qui sont les « mots » du langage.
Ces descripteurs peuvent être structurés sémantiquement et leur utilisation peut être régie par une syntaxe.

Une indexation contrôlée ne doit comporter que des éléments du LD, ce qui implique que ce langage soit exhaustif quant à la représentation des divers sujets traités dans le domaine couvert par la collection de documents.

Ce langage étant formel et figé, il comporte non seulement une liste de descripteurs, mais aussi des règles d'utilisation de ces descripteurs. Donc, il nécessite son apprentissage à la fois par l'indexeur et par l'interrogateur.

Pour avoir accès aux éléments du LD, un dispositif de présentation de ces descripteurs est nécessaire avec un classement logique [MANIEZ, 91]. Cette structure regroupe de proche en proche des notions plus fines sous des notions plus générales dont elles découlent avec un classement alphabétique des notions comme dans les dictionnaires encyclopédiques courants .

c➢ Indexation par un langage documentaire structuré (L.D.S.)

Le premier inconvénient relatif à la normalisation d'un LD est la nécessité pour l'indexeur et pour l'interrogateur de connaître le langage et ses règles d'emploi.
Pour remédier aux inconvénients des LD, le moyen de dictionnaires, de réseaux sémantiques [BALPE, 95] ou de tout processus permettant de passer d'un langage naturel à un langage cible par le biais :
- de l'indexeur, sous la forme d'une proposition de termes qui pourront être utilisés à l'interrogation,
- de l'interrogateur, sous la forme d'une proposition de termes qui ont pu être utilisés à l'indexation.

Ces propositions pourront être faites par l'étude des *relations sémantiques* répertoriées entre les mots du LD :

c.①. *Relation d'équivalence*

Cette relation est parfois citée comme faisant partie des éléments de structuration du LD (type thésaurus) et correspond à la rencontre d'une mention du type :

« mot_1 » : utiliser « mot_2 ».
Exemple: « États-Unis » : utiliser « USA ».

Il s'agit d'une relation liant un élément du langage naturel à un descripteur. Ce renvoie constitue une règle d'utilisation du descripteur appelé aussi le *synonyme préférentiel.*

c.②. *Relation d'ordre*

Les relations [terme générique/terme spécifique] et [tout/partie] sont des relations d'ordre partiel, qui permettent de construire des hiérarchies locales entre les descripteurs.

On peut représenter ces hiérarchies sous forme arborescente, et l 'ensemble de ces relations sur le LD constituent une polyhiérarchie (*Fig. II.3.1.c2.*).

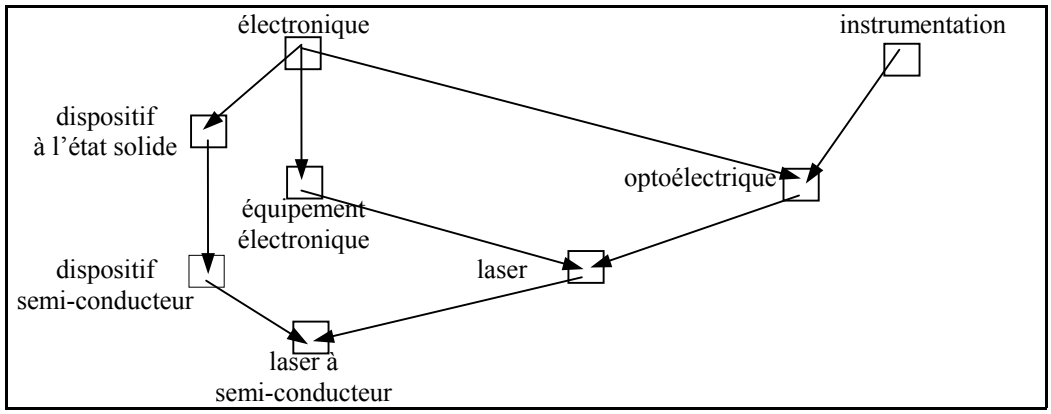


Fig.II.3.1.c2. Exemple de polyhiérarchie de descripteurs empruntée à DEWEZE A. 1981.

c.③. *Relation d'association*

Appelée aussi relation « voir aussi », cette relation permet de relier des descripteurs appartenant à des arborescences différentes. Elle est non réflexive, symétrique et non transitive.

c.④. *Relation de proximité*

Cette relation n'a de sens que si l'on a défini *un indice* de similarité, *un indice* de dissimilarité (distance entre les éléments du LD) et le choix d'*un seuil* pour la sélection de l'ensemble des descripteurs « proches » d'un descripteur donné.

La relation ainsi obtenue est réflexive, symétrique et non transitive.

Une formalisation de ces relations a été effectuée par A. Deweze [DEWEZE, 81], en transcrivant les descripteurs sous forme de « configurations sémiques » :

Une notion N_i est décrite par un sous-ensemble flou sur l'univers U ;
$U = \{s_i / i=1..n\}$ et s_i : le sème élémentaire constitué de l'ensemble des sèmes (S).

N_i est défini par :

1- l'ensemble des sèmes fixes de N_i pour lesquels : $\mu_{Ni}(S) = 1$

2- l'ensemble des sèmes latents pour lesquels : $0 < \mu_{Ni}(S) < 1$

3- l'ensemble des sèmes exclus de N_i pour lesquels : $\mu_{Ni}(S) = 0$

où : μ est une fonction d'appartenance sur l'intervalle de valeurs [0,1].

Remarques sur la structuration des descripteurs :

L'intérêt des relations qui viennent d'être présentées se situe principalement à deux niveaux :

♣ **Lors de l'interrogation**

Ces relations permettent de préciser une question lorsque celle-ci est trop générale et a fourni un trop grand nombre de documents :

- en utilisant la relation « *spécifique-générique »* pour particulariser un des éléments de la question ou se limiter à certaines des associations proposées par la relation « *voir aussi »* [BENMASSAOUD, 91] ;

Inversement, il est possible d'élargir une question :

- en remplaçant un descripteur par son terme générique, ou en explorant les descripteurs reliés par la relation « *voir aussi »* .

♣ **Lors de l'indexation**

Ces relations permettent à l'indexeur un élargissement du LD qui le rapprochera du langage naturel et sans nuire à la cohérence de l'indexation par :

- l'emploi de la synonymie et l'établissement des relations entre les descripteurs,
- la sélection d'un descripteur qui peut guider l'indexeur vers d'autres descripteurs intéressants [BEN ABDALLAH, 95], [BERTHOUZOZ, 97].

d➢ Indexation mixte (I.M.)

La combinaison des deux méthodes précédemment citées conduit à l'indexation mixte où l'indexation est effectuée à l'aide d'un LD et est complétée au moyen de descripteurs « libres ». Il s'agit surtout de deux indexations indépendantes qui coexistent pour le même document [GASTALDY, 92-95].

II.3.2- Indexation par structures

L'indexation qui serait constituée d'une liste libre de descripteurs s'avère insuffisante.
Par exemple, une indexation composée de termes « alimentation » et « poissons » peut faire référence aussi bien à des articles parlant de l'« alimentation des poissons » que d'autres traitant de l'«alimentation à base de poissons ». C'est pourquoi le LD peut être enrichi d'une syntaxe évoluée et compléter l'indexation.

Une étude du rôle de la syntaxe dans le LD réalisée par J. Maniez [MANIEZ, 77], a permis d'identifier les LD du point de vue de leur grammaire (*Fig.II.3.2.*) :

Indice (i)	Grammaire du LD
i = 0	Langage sans syntaxe.
i = 1	Langage indiquant simplement la présence de relations syntaxiques non spécifiées.
i = 2..n où n est un entier fini	Langage prévoyant une différenciation des liaisons syntaxiques en 2..n catégories.

Fig.II.3.2. Grammaire d'un LD selon J. MANIEZ 1977.

A titre d'exemple : soit la notion d'indice i d'une grammaire d'un LD et une syntaxe comprenant trois relations (R_j / j = 1..3). Nous pouvons déterminer plusieurs types de relations :

[R_1] : Rapports d'ordre formel ou intrinsèque
aucune relation logique ou physique entre les notions concernées.
***Exemple*:** « notion (a) » est à gauche de « notion (b) ».

[R_2] : Liaisons d'ordre intrinsèque de type statique
relations élément-propriété, agent-action, action-objet, etc.
***Exemple*:** « eau » - « liquide »,
« chaleur » - « oxydation »,
« oxydation » - « oxygène », etc.

[R_3] : Liaisons d'ordre intrinsèque de type dynamique
relation dite consécutive qui traduit une notion de conséquence, comme les relations agent-objet, facteur-produit, ...
***Exemple*:** « facteur » - « lettre »,
« lettre » - « enveloppe »/ « lettre » - « timbre », etc.

On constate que ces relations sont élaborées à partir de la lecture et la compréhension d'un texte en langage naturel :

- ♣ La relation [R_1] peut représenter la forme de surface de la phrase, soit la position des constituants de la phrase : l'analyse syntaxique.
- ♣ La relation [R_2] fait appel à la connaissance du lien sémantique entre divers éléments, soit l'élément et ses propriétés, l'action de l'objet, etc.
- ♣ La relation [R_3] fait intervenir la notion temporelle (logiques temporelles).

Remarques :

L'intérêt des diverses méthodes se mesure en fonction des performances réalisées, c'est-à-dire l'adéquation des documents fournis en réponse à une question donnée [CHAUMIER, 92].

Cette adéquation est le seul critère pour juger de l'efficacité de la méthode d'indexation. La mesure de l'adéquation fait appel à l'étude de quatre phénomènes qui sont des critères permettant de mesurer le résultat global de la requête. En outre ces critères font intervenir la notion de **pertinence** au travers plusieurs paramètres que sont la **précision,** le **rappel,** le **bruit** et le **silence** :

- La *précision* : elle mesure le taux de documents pertinents par rapport à l'ensemble des documents extraits :

$$\text{Pr}\acute{e}cision = \frac{nombre(documents\_pertinents\_extraits)}{nombre(documents\_extraits)}$$

- Le *rappel* : il mesure le taux de documents pertinents extraits par rapport à l'ensemble des documents pertinents contenus dans le corpus :

$$Rappel = \frac{nombre(documents\_pertinents\_extraits)}{nombre(documents\_pertinents)}$$

Plus la valeur de ces deux paramètres est grande plus le système de recherche est considéré performant. Nous pouvons également mesurer :

- Le *bruit* : il mesure le taux de documents non pertinents extraits par rapport à la totalité des documents extraits :

$$Bruit = \frac{nombre(documents\_NON\_pertinents\_extraits)}{nombre(documents\_extraits)}$$

- Le *silence* : il mesure le taux de documents pertinents non extraits par rapport à la totalité des documents pertinents contenus dans le corpus :

$$Silence = \frac{nombre(documents\_pertinents\_NON\_extraits)}{nombre(documents\_pertinents)}$$

La performance d'un système de recherche d'information peut être appréciée si, d'un côté, le taux de précision et de rappel est élevé et, de l'autre, le taux de bruit et de silence est bas.

Tel est l'objectif des méthodes d'indexation automatique que nous allons présenter dans la partie qui suit.

II.4- Indexation automatique

L'idée d'implémenter des systèmes d'indexation documentaires offre la possibilité de manier des objets virtuels (ou symboliques) sur ordinateur et sans encombrement physique. Ainsi, l'incontournable classement par auteur, et par titre sur les fiches de bibliothèque peut être avantageusement remplacé par l'accès aux documents à travers de multiples points d'entrée ou de critères [CHAUMIER, 90].

La recherche documentaire sur des champs fixes correspond aux fonctionnalités des bases de données classiques. Mais le système documentaire doit se montrer '*souple*' à l'interrogation de l'usager [BIÉBOW, 91] et '*intelligent*' pour répondre à sa demande :

> *« La solution est d'indexer les documents sur des indices ou concepts afin de modéliser le contenu, voire le sens .»* .

H.P. Luhn a suggéré, dès 1957, de se baser sur l'analyse de la fréquence des mots pour indexer les textes. En 1960, M.E. Maron et J.L. Kuhns introduisent la notion d'indexation

probabiliste et bâtissent le premier système d'indexation automatisée KWIC (Key Word In Context) [BLANQUET, 93-94].

Ces précurseurs sont suivis dans leur démarche par G. Salton, qui en 1966, montre la nécessité de se tourner vers l'indexation de textes. Une récente étude de G. Salton [SALTON, 86a], en 1986, comparant les performances entre l'indexation contrôlée et l'indexation automatique, a montré les potentialités de l'automatisation [SALTON, 86b].

II.4.1- Indexation automatique à base de mots-clés

Les systèmes bénéficient d'une automatisation de l'indexation basée sur l'observation des textes ou de leurs résumés. Les procédures se réduisent le plus souvent à des traitements très simples basés sur l'identification des mots-clés préalablement choisis. Les textes fournis à l'utilisateur sont ceux reliés aux mots-clés dans la formulation de sa question.

Une première automatisation peut être l'indexation sur les mots du titre des ouvrages ou des articles, à l'exemple du système AQUEDUCT en 1988 [ANTON, 88]. Mais, le contenu d'un texte est rarement défini par son titre.

Dans des systèmes plus élaborés, ces mots-clés sont déterminés automatiquement par l'élimination des mots fonctionnels (mots supposés « non informatifs » comme l'article, la préposition, l'auxiliaire, etc.) de l'ensemble des mots du corpus. Ceux qui restent après filtrage obtiennent le statut de mot-clés au sens du mot « informatif ».

Inconvénients :

- ***Normalisation*** : il n'y a pas de normalisation des mots. Les deux mots « système » et « systèmes » sont différemment perçus.
- ***Détection*** : les systèmes ne distinguent pas plus les syntagmes que les mots-composés, ce qui introduit inutilement du bruit. Les deux mots « apprentissage » et « informatique » peuvent se retrouver dans des configurations de sens différents comme « apprentissage de l'informatique », « informatique et apprentissage de la lecture », « les méthodes d'apprentissage en informatique », etc.. La recherche sur ces index conduit inévitablement à rappeler un nombre de textes hors sujet.
- ***Synonymie et homonymie*** : les synonymes et les homonymes ne sont pas distingués.
- ***Ambiguïté*** : les ambiguïtés morphologiques des mots ne sont pas résolues. Dans le cas du français, 30% en moyenne des mots pris isolément sont ambigus.

L'amélioration des systèmes d'indexation automatiques passe par la définition des procédures plus rentables que les mots-clés et la sélection de termes de manière plus fine rend l'indexation plus efficace.

II.4.2- Indexation automatique à base d'outils statistiques

Les outils statistiques tentent de parvenir à une représentation du sens des textes qui les caractérisent les uns par rapport aux autres. La représentation d'un document est calculée en fonction de l'ensemble des documents contenus dans la base. L'idée suggère de se baser sur l'analyse de la fréquence des mots pour indexer les documents (ou textes) [BORIAS, 95], [CHEN, 98a].

Ces outils présentent la particularité de ne tenir aucun compte de la sémantique des mots et le seul compte est le calcul de différents indices mathématiques et statistiques.

II.4.2.1- Mesure du pouvoir discriminant d'un terme

L'idée de « discrimination », dans le cadre des recherches sur les bases de données documentaires, consiste à calculer le poids informationnel de chaque item par une fonction destinée à faire émerger une distinction entre les documents [YU, 77].

Cette fonction est à la base de la théorie de l'information :

- Plus un mot est rare sur l'ensemble des documents et fréquent dans un document particulier, plus il est caractéristique du document.
- Le pouvoir d'un terme i pour représenter un document j est basé sur sa fréquence dans le document « tf_{ij} » (*term frequency*) et sur l'inverse du nombre de documents de la base « idf_{ij} » dans lesquels il apparaît (*inverse document frequency*).

G. Salton propose un premier modèle [SALTON 86b] très simple : $W_{ij} = tf_{ij} * idf_{ij}$

Le terme idéal (W_{ij}) est celui qui apparaît beaucoup dans un document tout en étant particulièrement rare dans les autres.

Un modèle probabiliste plus élaboré [SALTON 86b] consiste à mesurer (tr_{ij}) *l'intérêt d'un terme i* : $tr_i = \log \frac{N - idf_i}{idf_i} + c$

avec :

N : nombre total de documents de la base,

idf_i : nombre de documents dans lesquels apparaît le terme i,

c : une constante.

II.4.2.2- Mesure basée sur la densité de l'espace des documents

Soit une base documentaire dans laquelle les documents sont liés à des index. La représentation de ces liens se fait sous la forme d'un vecteur d'indexation.

Le calcul des mesures de similarité entre les vecteurs traduit les similarités entre les documents. Un bon index se traduit par son pouvoir séparateur entre les documents qu'il rend moins similaires les uns aux autres et l'introduction d'un tel index fait diminuer la densité de l'espace [SALTON, 73].

Dans ce cas, la valeur de discrimination d'un terme appelée DV (*Discrimination Value*) est égale à la différence de densité de l'espace avant et après indexation de ce terme.

La difficulté est de définir la densité de l'espace puisque celui-ci n'est pas euclidien. Une solution consiste à assimiler cette densité à la moyenne des similarités entre les documents pris deux à deux. Une autre solution consiste à définir un document centroïde puis calculer la moyenne des similarités entre chaque document à celui-ci.

Cette mesure permet de ranger les termes en fonction de leur pouvoir discriminant par ordre décroissant et de distinguer les termes problématiques dans le but d'un futur ajustement.
Un terme problématique est celui dont le DV est fortement négatif ou proche de zéro.

II.4.2.3- Analyse de la sémantique latente

(en anglais : *Latent Semantic Analysis* [LANDAUER, 91]).

Lorsqu'un document n'est pas retrouvé alors qu'il répond à une demande, dans ce cas, les mots de l'interrogation ne sont pas identiques à ceux du texte, mais ils sont des synonymes. Dans le cas contraire, les homonymes provoquent le rappel de textes non pertinents.

Les auteurs de cette méthode statistique supposent que les textes sont porteurs d'une structure sémantique implicite dont ils tentent d'extraire les concepts en tant qu'unité de sens.

La méthode utilise comme données une matrice représentant les documents sur les colonnes et les termes sur les lignes. Pour la ligne i et la colonne j, la valeur représentée est la fréquence du terme i dans le texte j.

La technique de décomposition en valeurs propres (en anglais : *singular value decomposition*) permet de réduire cette matrice dans un espace réduit de dimensions orthogonales. Le nombre de dimensions orthogonales est adéquat à la représentation du domaine fixé.
L'originalité de la méthode est de réduire les dimensions de l'espace en modélisant les variations sémantiques significatives tout en diminuant le bruit.

Soit *Y*, la matrice représentant les termes de tous les documents. *Y*, matrice rectangulaire, peut s'écrire comme le produit de deux matrices carrées et ayant l'avantage d'avoir les mêmes valeurs propres non nulles : $Y_{t,d} = Y_{t,t}^{T} * Y_{d,d}$

Y peut être décomposée en trois matrices :

1- ***Doc***, matrice des documents,
2- ***Term***, matrice des termes,
3- ***Diag***, matrice diagonale des valeurs propres.

La formule de reconstitution de Y est : $Y_{t,d} = Term_{t,t} * Diag_{m,m} * Doc_{m,d}$

avec : *m*, dimension de la matrice;
t, nombre de termes;
d, nombre de documents.

Les valeurs propres de la matrice *Diag* sont rangées suivant leur pertinence décroissante. De cette matrice, les plus grandes valeurs propres sont gardées jusqu'au rang k pour former la matrice ***Diag'***.
Les matrices *Term* et *Doc* sont transformées. Seules les colonnes et les lignes correspondant aux valeurs propres de *Diag'* sont conservées pour former les matrices ***Term'*** et ***Doc'***.

Soit *Y'* : $Y'_{t,d} = Term'_{t,k} * Diag'_{k,k} * Doc'^{T}_{k,d}$

Y' est l'unique matrice de rang k la plus proche de Y au sens des moindres carrés.

Les termes utilisés pour l'indexation peuvent être de simples mots ou de syntagmes nominaux déterminés par une procédure semi-automatique. Après la décomposition, les documents et les termes sont regroupés en collections et en fonction de leur proximité spatiale [DIDAY, 2001].

II.4.2.4- Modèles de distribution statistique

Cette approche fait l'hypothèse que les index sont distribués sur les documents suivant des lois statistiques [BOOKSTEIN, 75].

Sur un ensemble de documents de même taille, on s'attend à ce que :

- Les occurrences des mots porteurs d'un sens spécifique (mots sémantiques) soient regroupées dans un petit nombre de documents et non distribuées uniformément.
- Les occurrences des mots non spécifiques (articles, prépositions,...) soient aléatoirement distribuées.

La propriété de regroupement des mots sémantiques dans un petit nombre de documents est utilisée pour distinguer les futurs index.
Les documents sont considérés homogènes en regard des mots distribués de façon aléatoire. De la même façon, un mot (respectivement, une classe de mots) d'indexation qui caractérise un ensemble de documents est distribué de façon aléatoire par rapport à cet ensemble.

Soient les documents de la classe i.
Soit λ_i, le nombre de mots attendus dans un document de la classe i.
La probabilité qu'il y ait k occurrences d'un mot donné associé à la classe i est donnée par la formule suivante : $\Pr\{k / i\} = \frac{\lambda_i^{\,k}}{k!} e^{-\lambda_i}$

Mais, la classe d'un document n'est pas connue :

Soit π_i , la probabilité qu'un document appartient à la classe i : $P(i) = \pi_i$.
La probabilité qu'un document quelconque contienne k occurrences d'un mot donné : $P(k) = \sum_i \pi_i \frac{\lambda_i^{\,k}}{k!} e^{-\lambda_i}$

Si on connaît la fréquence k du mot (w) dans un document donné, alors la probabilité que ce document appartienne à la classe i (théorème de Bayes) :

$$\Pr\{i / k\} = \frac{P(i)}{P(k)} \Pr\{k / i\}$$

Soit r_i, la probabilité qu'un document de la classe i soit jugé pertinent.

si une personne demande des documents à propos du mot (w) et qu'un document contient k occurrences de w,
alors la probabilité que le document soit jugé pertinent est :

$$\Pr_w(k) = \sum_i r_i \Pr\{i / k\} \frac{\lambda_i^{\,k}}{k!} e^{-\lambda_i} = \frac{\sum_i r_i \pi_i \lambda_i^{\,k} e^{-\lambda_i}}{\sum_i \pi_i \lambda_i^{\,k} e^{-\lambda_i}}$$

D'autres auteurs ont poursuivi cette approche, notamment A. Andreewsky, qui généralise cette approche en étudiant les caractéristiques discriminantes d'un document en fonction de l'ensemble de symptômes mesurés [ANDREEWSKY, 76].

Un symptôme se traduit par l'identification d'un mot, d'un syntagme, ou toute expression jugée appropriée.

Il conclut que l'efficacité d'un discriminant est obtenue lorsque les fonctions de poids qui affectent les symptômes sont bien choisies.

La différence de comportement des mots face aux modèles mathématiques est un critère d'indexation, comme le théorème de Bayes qui permet de distinguer deux classes de termes : celle qui est proche de la loi envisagée pour la distribution des index et l'autre.

D'autres recherches ont abouti au choix du modèle de Poisson. Les chercheurs l'ont qualifié de modèle plus adapté à la distribution de termes dans les textes de façon à discriminer les items d'indexation, car les mots fonctionnels semblent être distribués suivant cette loi.

Le cas du modèle de distribution de S.P. Harter correspond à une loi de 2-Poisson [HARTER, 75] :

$$f(k) = \pi \frac{\lambda_1^{\ k}}{k!} e^{-\lambda_1} + (1-\pi) \frac{\lambda_2^{\ k}}{k!} e^{-\lambda_2}$$

S.P. Harter évalue les paramètres ($\pi, \lambda_1, \lambda_2$) en ajustant les distributions à ses données par la méthode des moments. Il suppose que les documents sont divisés en deux classes : ceux concernant le sujet en question et les autres qui y sont périphériques.

L'amélioration de ce modèle suggère une extension des classes : trois classes sont définies et concernent les documents centrés sur le sujet, les documents périphériques et les documents externes au sujet.
Les calculs dans le modèle Poissonien présentent l'inconvénient de demander des temps de calculs qui croissent exponentiellement avec la taille de la base.

Remarques :

Le problème majeur de tous ces modèles et approches statistiques est que dans un modèle statistique les fonctions discriminantes ne parviennent pas à maîtriser le pouvoir d'expression des mots significatifs : les taux de bruit et de silence restent relativement importants selon la taille des corpus traités [EL GUEDJ, 97]. L'amélioration de certaines fonctions de distribution rèste limitée [KODRATOFF, 96].

II.4.3- Indexation automatique à base d'outils linguistiques

Les raisons qui ont poussé les spécialistes de la documentation à s'intéresser aux théories linguistiques sont nombreuses. Les propriétés des langages d'indexation (LI) ressemblent beaucoup à celles des langages naturels (LN) et certaines en dérivent profondément.

Comme les LI sont appliqués principalement à des textes exprimés en LN, le problème de passage d'un langage à l'autre se pose :

- **J. Maniez** [1980] soulève dans sa thèse une question générale : quelles sont les différences et les ressemblances nécessaires entre LN et LI ? Pour lui, l'élément commun est l'universalité de la fonction référentielle (la nature symbolique de signe). L'usager ne cherche pas de termes d'indexation pour eux-mêmes mais pour les documents dont ils représentent le sujet [MANIEZ, 93].
- **J-C. Gardin** [1973] qualifie les LI par rapport aux LN de métalangage (langage ou système de symboles) pour exprimer le contenu d'un document rédigé en LN [GARDIN, 74].
- **R. Fugmann** [1982] considère les LI et les LN comme complémentaires. Les concepts individuels référant à un seul objet sont exprimés en LN par une seule expression lexicale. Par contre, les concepts généraux référant à une multitude d'objets sont souvent rendus en LN par plusieurs expressions lexicales (synonymes), voire par des expressions non lexicales (périphrases). Ce constat se porte dans le domaine scientifique où la terminologie est en retard sur les notions [FUGMANN, 83-93].

II.4.3.1- Les universaux du langage

Dans les années soixante, la linguistique structuraliste (N. Chomsky, C.J. Fillmore, B. Pottiers, J. Lyons, etc.) a cherché à modéliser une structure profonde qui permet de rendre compte des structures de surfaces aussi diverses que sont les langues naturelles.

Dans le domaine des sciences de l'information, on peut se demander *si les modèles linguistiques ne seraient pas suffisamment fondamentaux pour expliquer aussi la structure des langages artificiels (ou métalangage) comme les LI*.

Par exemple, la grammaire des cas de C.J. Fillmore permet de fournir une liste de catégories fondamentales utilisables pour la syntaxe des LI [MANIEZ, 77].

Une comparaison de la structure profonde d'un texte à la liste de ses index montre bien que cette structure se situe déjà à un niveau de généralité plus élevé [SMEATON, 89-91a].

II.4.3.2- Aspects linguistiques

La partie linguistique des systèmes de traitement de la langue naturelle qui nous intéressent assure la reconnaissance des éléments du texte en fonction des niveaux d'analyses ci-dessous :

- L'analyse morphologique : reconnaissance des lexèmes dans le lexique et la normalisation des mots fléchis.
- L'analyse syntaxique : construction de la représentation syntaxique. Les différents groupes de la phrase sont délimités ainsi que les relations entre ces groupes.
- L'interprétation sémantique : construction d'une représentation sémantique à partir de la représentation syntaxique précédente et la prise en compte du contexte.

Ces modules permettent de désambiguïser les mots et d'identifier les synonymes et les relations hiérarchiques. L'organisation de ces modules n'est pas obligatoirement respectée et les tâches syntaxiques et sémantiques sont parfois effectuées en parallèle, comme dans la *grammaire de Montague* [WARREN, 82] ou les *grammaires syntagmatiques généralisées*.

II.4.4- Les outils linguistiques

II.4.4.1-Introduction

Tout **S**ystème d'**I**nformation **D**ocumentaire (S.I.D.) se caractérise par une fonction d'indexation et une fonction d'interrogation (ou de recherche), qui à partir d'un corpus de documents en permettent la recherche d'informations [BERRUT, 86-89].

L'adéquation entre requête et documents est évaluée par rapport à une fonction de correspondance, qui est elle-même fondée sur un modèle abstrait "modèle de correspondance".

L'idée d'introduire des outils linguistiques, dans les SID, est une approche comme d'autres pour apporter certaines améliorations sur la qualité de ces systèmes sur plusieurs niveaux d'analyse :

- ♣ ***Niveau interface Homme-Machine***
  La présentation de l'interface de dialogue avec l'utilisateur au moment de l'interrogation : un dialogue en langage naturel [SYSTEX, 92].
- ♣ ***Niveau de la définition des termes d'indexation***
  Les termes peuvent être envisagés comme des unités linguistiques complexes, telles que par exemple des syntagmes nominaux simples ou complexes, au lieu des traditionnels termes.
- ♣ ***Niveau des réponses données à l'utilisateur :***
  Au lieu que les réponses soient de simples références aux documents, elles pourraient être explicites en présentant le résumé du document [SMEATON, 91b], sa table de matière, etc., au travers d'un dialogue homme-machine.
- ♣ ***Niveau de la génération :***
  Des fonctions de transformation de textes, tels que des générations de résumé ou bien des procédures de traduction partielle [TOUSSAINT, 97], [DENOUE, 2000].
  etc.

Tous ces niveaux sont complexes par la nature de la langue elle-même. Pour être réaliste, tout traitement automatique de la langue naturelle ne peut être envisagé que pour atteindre des objectifs bien délimités [WOODS, 98a-b], de manière à pouvoir développer les outils nécessaires et adéquats.

Classiquement, tout traitement de la langue naturelle peut se décomposer an quatre niveaux : la morphologie, la syntaxe, la sémantique et la pragmatique. Pour être complet, y ajouter la phonologie et la lexicologie.

Chaque traitement présente, en fonction des objectifs qui lui sont assignés, un degré d'intervention de profondeur propre pour chacun de ces niveaux. Il se peut par ailleurs qu'un ou plusieurs de ces niveaux soient inexistants (pour un traitement donné) ou privilégier un niveau sur un autre dans le processus [ABEILLÉ, 95].

En général, quel que soit le domaine traité, cette discussion porte sur la profondeur respective demandée aux niveaux syntaxique et sémantique.

Cette discussion a donné naissance à deux grands courants de modèles linguistiques en système d'information :

- le premier prône l'utilisation de la syntaxe dans le but d'extraire des structures syntaxiques et représentatives a priori du contenu d'un document,
- le second courant met en exergue une sémantique forte permettant une représentation relativement fine des documents.

II.4.4.2- Les modèles à dominante syntaxique

II.4.4.2.1- Grammaire Générative Transformationnelle

La grammaire générative transformationnelle (G.G.T.) de la fin des années soixante-dix est le modèle dominant à l'époque sous le nom de Théorie Standard Etendue. N. Chomsky (1965) est pris pour un des tenants de cette théorie [CHOMSKY, 86-87].

D'un point de vue épistémologique, la GGT opère un déplacement de l'empirisme vers le rationalisme avec comme critère le « ***rasoir d'Occam*** », c'est par la proposition suivante : *« Ne pas multiplier les suppositions au-delà du minimum nécessaire ».* Il s'agit d'un critère de simplicité à plusieurs niveaux [TCHOBANOV, 95] :

- ***Simplicité Descriptive***
  On retient parmi plusieurs grammaires, qui rendent compte des faits linguistiques, la plus simple et la plus compacte.

- ***Simplicité Inter-linguistique***
  Les différences entre les « bonnes » grammaires des différentes langues doivent être minimales. Une théorie générale des grammaires traite de la prédisposition des humains aux langages (innéisme).
  Ceci est motivé par l'observation suivante : *« le langage humain est un choix trop particulier pour être découvert empiriquement dans l'infinité des langages formellement possibles ».*

- ***Simplicité Inter-disciplinaire***
  Si une théorie linguistique confirme mieux les hypothèses d'une bonne théorie psychologique qu'une autre à couverture égale, alors la première théorie sera retenue.

Ce dernier critère est au fondement de la doctrine des sémanticiens génératives, comme mouvement dans le modèle GGT datant de la fin des années soixante et début des années soixante-dix (Seuren, Katz, etc.). Ils considèrent la sémantique comme une science interdisciplinaire réunissant le efforts des logiciens, linguistes et psychologues à l'instar des sciences cognitives modernes.

Du point de vue de la hiérarchie de N. Chomsky et M.P. Schützenberger (1970), une GGT est une grammaire mixte. Elle possède une composante de règles indépendantes du contexte, qui produit des ***structures profondes*** servant d'entrée à des règles transformationnelles non-contraintes (type 0) [SCHÜTZENBERGER, 77], qui produisent des ***structures de surfaces*** [ABEILLÉ, 98] (*Fig. II.4.4.2.1.*)**.**

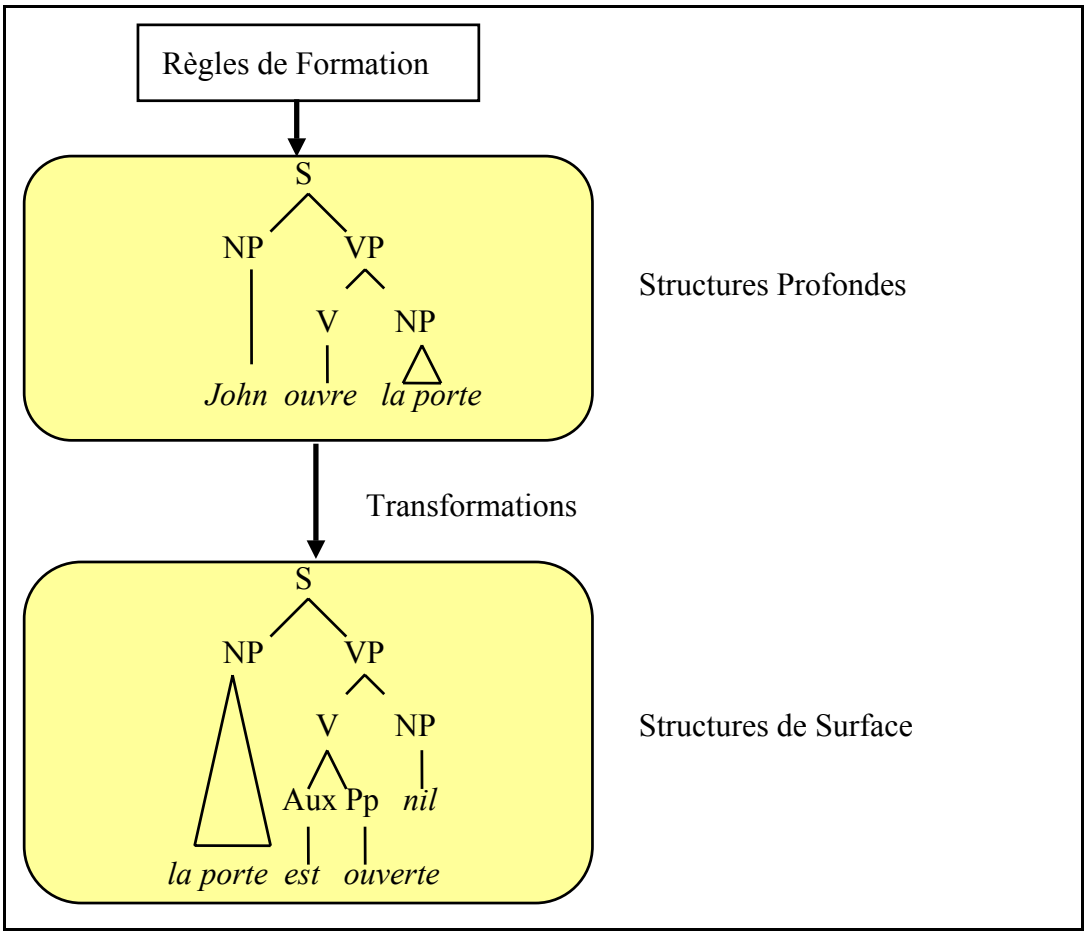


Fig. II.4.4.2.1. GGT Modèle Classique .

Remarques :

♣ ***Place de la sémantique dans ce modèle***

Un premier débat opposait, d'une part, les tenants d'une sémantique générative s'appuyant sur une structure profonde de plus en plus abstraite pour représenter toutes relations sémantiques, et d'autre part, les partisans d'une sémantique interprétative s'appuyant davantage sur les structures de surfaces et aboutissant à la notion de forme logique pour la représentation du sens.

Les seconds développaient, en liaison avec des logiciens, des modèles sémantiques formalisés dans la lignée des travaux de R. Montague pour un calcul mathématique du sens.

♣ ***Plausibilité de l'appareil transformationnel***

Les travaux effectués en mathématique linguistique, qui montraient que la classe des GGT était formellement équivalente à celle des grammaires non-contraintes. L'appareil théorique employé ne permettait pas de caractériser formellement la classe des langues naturelles (humaines) dans l'ensemble de tous les langages abstraits reconnaissables par la machine de Turing.

♣ ***Adaptation à la construction de programmes***

Le modèle transformationnel a révélé la difficulté de son adaptation à la construction de programmes informatiques pour l'analyse automatique de textes.

II.4.4.2.2- Exemple de modèle : IOTA

Le système que nous présentons ici est le prototype IOTA développé à Grenoble [CHIARAMELLA 86]. La base de l'indexation de IOTA repose sur l'hypothèse que l'information véhiculée dans les textes se retrouve essentiellement dans les syntagmes nominaux. L'indexation dans IOTA se divise en deux phases :

- ♣ ***une phase d'extraction des syntagmes nominaux***
  Dans cette phase, l'extraction est basée sur un analyseur de surface de la langue naturelle entièrement automatique avec la possibilité d'enrichir le lexique de manière automatique lors de la rencontre de mots nouveaux.
- ♣ ***une phase de génération et de pondération des termes d'indexation à partir des SN extraits d'un thesaurus prédéfini***
  La génération des termes d'indexation est fondée sur un processus de transformation des groupes nominaux reconnus dans le texte et une comparaison avec les éléments d'un thésaurus.

Le thesaurus IOTA est utilisé tant à l'indexation qu'à l'interrogation et il est construit par un processus entièrement automatisé qui est fondé sur des critères syntaxiques et statistiques :

une évaluation des liaisons contextuelles entre certaines classes de termes dans un corpus représentatif permet d'enregistrer certaines mesures dans une matrice de termes. L'exploitation de cette matrice permet d'en extraire des sous-graphes maximaux et complets, appelés des ***cliques***, qui représentent des classes de concepts.
Une clique représente l'ensemble des concepts correspondants aux groupes nominaux qui peuvent être construits à partir de ses constituants.

L'interrogation dans IOTA est bâtie autour d'un système expert permettant une interrogation en langage quasi-naturel :

la modélisation « typologie » de l'usager, l'évaluation des références résultat de la question et la reformulation automatique permettent à l'utilisateur de remodeler sa requête après une évaluation jugée insuffisante.

II.4.4.3- Les modèles à dominante sémantique

II.4.4.3.1- Sémantique Générative et Grammaires d'Unification

♣ ***Grammaires d'Unification :***

On peut considérer que les grammaires d'unification (G.U.) sont des grammaires à base de contraintes, comme les nouvelles théories syntaxiques qui datent des années quatre-vingts.

Il s'agit de modèles qui recherchent une articulation explicite entre lexique, syntaxe et sémantique. Les propriétés linguistiques correspondantes sont conçues comme des informations associées aux morphèmes, aux syntagmes ou aux constructions. Ces propriétés sont combinées par des opérations variées dont l'unification occupe une place centrale [CHARDENON, 97], [ABEILLÉ, 98].

Dans cette conception des GU, des modèles logiques ou mathématiques sont introduits, pour lesquels des méthodes de programmation sont définies. En général, un compromis entre

l'expressivité linguistique et l'efficacité du code informatique [CHAUDIRON, 94] incorpore les concepts linguistiques.

Le développement des GU se caractérise par une importation sans précédent d'outils formels développés dans un domaine varié, principalement en logique et linguistique computationnelle. La construction d'outils et de programmes pour la traduction automatique, l'aide à la correction orthographique, l'indexation automatique ont également contribué au développement de nouveaux modèles syntaxiques.

L'approche alternative, celle d'une sémantique et d'une Grammaire de Surface est due aux travaux du logicien R. Montague. Ce dernier considère que les structures de surface (langage formel) détiennent une valeur sémantique d'une expression calculée à partir de certaines valeurs assignées aux éléments qui la constituent [WARREN, 82]. Cette approche connue comme « sémantique formelle » a inspiré bon nombre de théories linguistiques modernes dans le cadre des GU comme les formalismes d'unification : GPSG (*Generalized Phrase Structure Grammar*), LFG (*Lexical Functionnal Grammar*), HPSG (*Head-driven Phrase Structure Grammar*), etc.

Dans ces théories, une phrase bien formée est considérée comme un calcul de vérité réussi sur les traits sémantiques, syntaxiques et phonologiques assignés aux items lexicaux. En plus, ces approches proposent un traitement sans transformations des phénomènes ayant servi d'arguments majeurs à la théorie transformationnelle (tels le passifs, les dépendances à distance, etc.).

♣ *Sémantique Générative :*

L'approche novatrice de Katz et Postal (1964-70) fut largement reconnue comme valable, même si la notion de Représentation Sémantique (RS) manquait de définition précise.
Au cours de la recherche de fondements empiriques pour les RS, la découverte des phénomènes mettant en cause la différenciation entre des Structures Profondes (SP) et des RS s'est établie (*Fig. II.4.4.1.3.1a*).

L'argumentation reposait entre autre sur un grand nombre de cas, où des règles qui génèrent les Structures de Surfaces (S.S.) correctes (et éliminations des SS agrammaticales) jouent un rôle crucial pour la bonne interprétation sémantique des phrases.

La proposition des sémanticiens générativistes était de faire l'économie des SP et de considérer les RS comme des structures linguistiques arborescentes qui forment l' « input » de la composante transformationnelle (*Fig. II.4.4.1.3.1b.*). Ces structures seraient formulées dans un langage formel et logique suffisamment abstrait pour rendre sémantiquement interprétable les règles de projection. Le niveau des structures profondes syntaxiques deviennent donc inutiles.

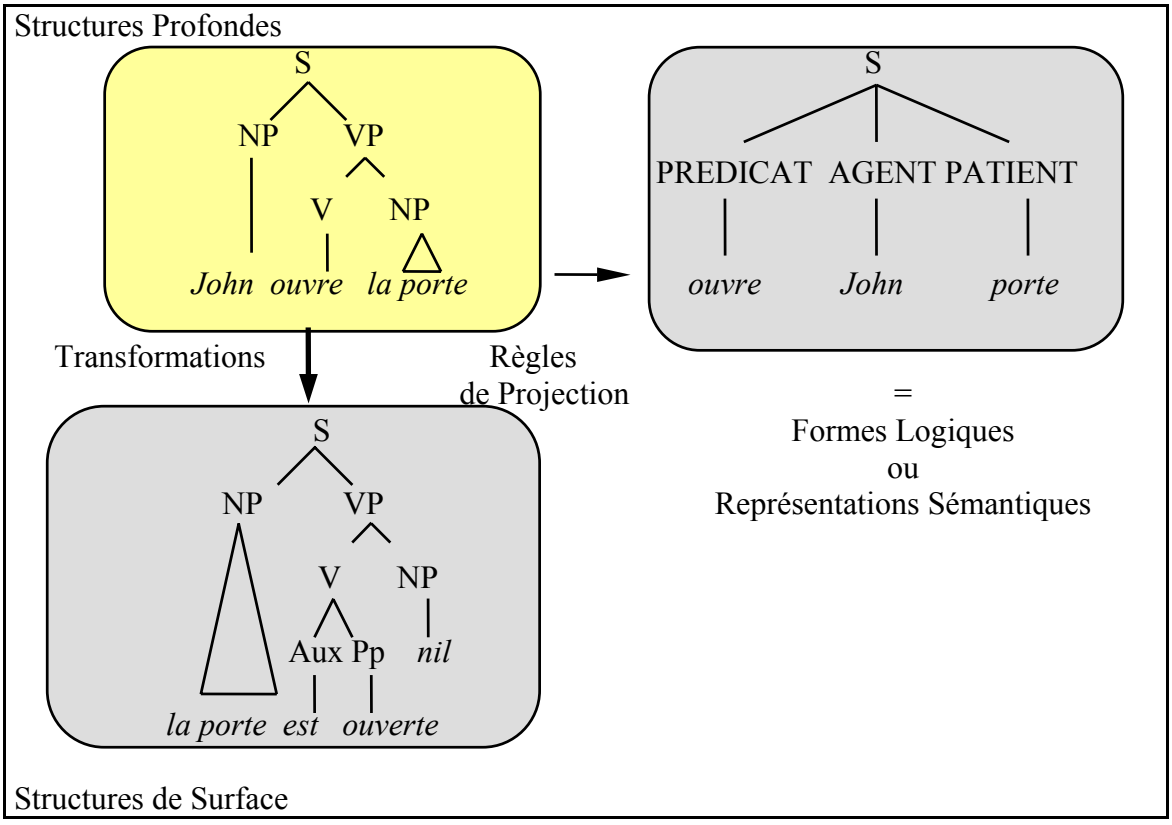


Fig. II.4.4.1.3.1a. GGT selon Katz et Postal, 1964.

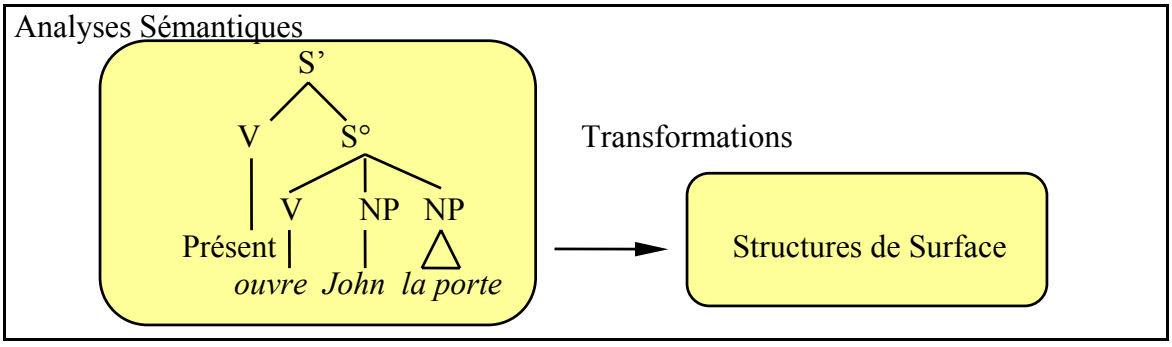


Fig. II.4.4.1.3.1b. GGT selon la Sémantique Générative.

Critiques de la Sémantique Générative :

♣ Les tenants d'une syntaxe autonome adressent de sévères critiques à ce mouvement naissant : l'écriture « indisciplinée » des règles et le manque de contraintes dans l'édifice formel.

♣ L'étude de Contraintes Universelles, ce qui revenait à admettre les défauts de leurs propres modèles (GGT) : la critique de l'insuffisance des contraintes était détournée vers la sémantique générative.

♣ L'évolution de la syntaxe autonome vers des modèles sans transformations : la restitution de nouveaux modèles sur le marché actuel des théories linguistiques.

II.4.4.3.2- Exemple de modèle : RIME

Le système RIME [BERRUT, 88] propose d'étudier des textes en langue naturelle. Il les représente dans un modèle sémantique relativement souple. Il intègre des procédures autonomes traitant de la morphologie, de la syntaxe et de la sémantique. Ces modules sont écrits dans un langage déclaratif Prolog incorporant plusieurs fonctionnalités d'une grammaire d'unification :

♣ ***Intervention des processus***
chacun des processus a été fixé à partir d'une étude des phénomènes linguistiques. Ce qui est une condition nécessaire à une bonne mise en oeuvre du traitement linguistique.

♣ ***Processus de coopération***
un processus de coopération réalise l'ordonnancement et l'indépendance des processus linguistiques.

Les documents traités sont des comptes rendus médicaux décrivant des images radiologiques. Une partie essentielle fait par RIME [BERRUT, 89] consiste à traiter les parties textuelles de ces comptes rendus en les traduisant selon un format interne de représentation des connaissances ayant pour but de définir :

- un format de représentation, appelé le modèle sémantique
- la représentation des phénomènes linguistiques, pour permettre la traduction de textes en langue naturelle dans ce modèle sémantique.

L'idée de base dans le modèle sémantique de RIME consiste à représenter le sens de chaque phrase par un graphe sémantique de type arborescent.

Une fois les comptes rendus médicaux structurés, leurs contenus sémantiques deviennent facilement représentable selon des principes de dépendance conceptuelle. Les connaissances véhiculées dans ces documents seront donc représentées au travers de schémas mettant en évidence des relations sémantiques entre différents concepts intervenants dans les textes traités. Les concepts et les relations sémantiques sont représentés comme suit :

A- Structures d'arbres binaires complets

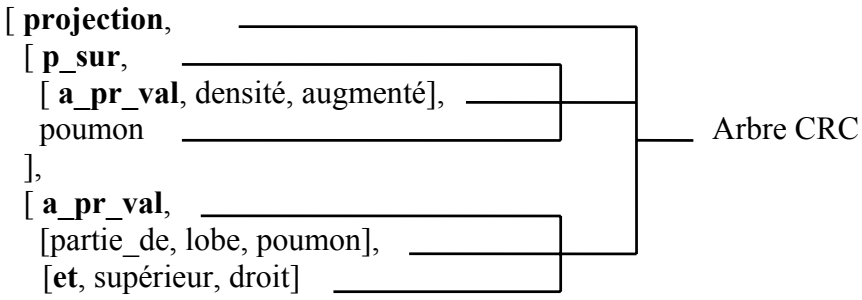
Les concepts et les opérateurs sémantiques se représentent par des structures d'arbres binaires complets dans lesquelles :

- *les noeuds correspondent aux opérateurs sémantiques*
  Ces opérateurs sont binaires et ils explicitent le lien sémantique entre les concepts représentés par les sous-arbres gauche et droit de l'arborescence.
  Une représentation des opérateurs sémantiques dans la grammaire du modèle RIME est comme suit : [**a_pr_val**, volume, augmenté].
  A travers cette représentation, elle indique une arborescence binaire complète en indiquant le parcours préfixé, ie. [**père**, fils_gauche, fils_droit].
- *les feuilles correspondent à des faits médicaux ou à des termes techniques*
  A ce niveau, les faits médicaux ou les termes techniques relatifs à un examen médical effectué correspondent au vocabulaire de base du modèle.
  Par exemple, poumon, cancer, opacité, etc.

B- Réduction en arbres : Compte rendu conceptuel

Chaque phrase d'un compte rendu médical est traduite en une arborescence (cf. a-) et l'ensemble de ces arborescences représente le « sens » du compte rendu médical, appelé *compte rendu conceptuel* (C.R.C.).

Par exemple :
la phrase « condensation pulmonaire en projection du lobe supérieur droit » se traduit comme suit :



C- Langage conceptuel

Les arbres construits doivent respecter le modèle formel qui est défini par une grammaire. Le langage engendré par cette grammaire s'appelle le « langage conceptuel ».

L'organisation interne des comptes rendus médicaux peut se décrire au travers d'une grammaire dans laquelle les méta-symboles correspondent à des concepts intermédiaires. Ces derniers sont représentés par les signes, les lésions, les constatations et les symboles terminaux représentant des opérateurs sémantiques, des concepts atomiques ou des fonctions de l'organisme humain.

Il existe trois niveaux fondamentaux dans l'organisation des comptes rendus médicaux (C.R.).

♣ ***Niveau 1***

Ce niveau exprime qu'un Compte rendu médical est constitué d'une ou plusieurs phrases permettant de relier plusieurs composants d'un compte-rendu.

Par exemple :
CR::= CONSTAT
CR::= DIAGNOSTIC
CR::= [**permet_de_déduire**, CONSTAT, DIAGNOSTIC]

Ces règles permettent de définir la grammaire du CR, qui est constitué d'un seul constat, d'un seul diagnostic ou bien des deux. Le constat permet de déduire un diagnostic. Le lien sémantique entre ces deux derniers se traduit par l'opérateur sémantique **permet_de_déduire**.

♣ ***Niveau 2***

Le second niveau définit les notions de constat, les notions de diagnostic et les sous-notions qui leur sont associées.

Par exemple :

CONSTAT::=[dû_à, CONSTAT, CONSTAT]	*exprime que des constats peuvent être indépendants.*
CONSTAT::=[**montre_par**,SIGNE,CONSTAT]	*ces règles définissent un constat à partir d'un signe*

	seul, ou bien par un signe révélé par un examen. *Un signe est une entité observable, telle une cavité.*
DIAGNOSTIC::=[**et**,DIAGNOSTIC,DIAGNOSTIC] etc.	*un diagnostic comme une combinaison de plusieurs diagnostics.*

♣ ***Niveau 3***

Le troisième niveau contient les règles pré-terminales et les règles terminales de la grammaire, et correspond aux concepts de plus bas niveau du modèle RIME. L'ensemble des symboles terminaux, hormis les opérateurs sémantiques, est consigné dans un lexique. A chaque entrée lexicale une information sémantique est attribuée.

Par exemple :

SIGNE::= {t ∈ V_T / catégorie_sémantique(t)='signe'},

LESION::= {t ∈ V_T / catégorie_sémantique(t)='lésion'}, etc.

Cette grammaire définit donc un langage conceptuel de RIME dont l'utilisation permet une interrogation fine.

II.4.4.4- Conclusion

Tous ces systèmes à dominante syntaxique ou sémantique présentent des aspects non seulement intéressants, mais également des imperfections souvent liées à des procédures linguistiques *insuffisantes*, *lourdes* ou *complexes à mettre en oeuvre* :

- ***Insuffisantes***

ces procédures ne gèrent que certains phénomènes linguistiques, le reste pouvant constituer des états de blocage du système ou des résultats sans fondements.

- ***Lourdes***

le partage des tâches, entre les différents processus linguistiques potentiels, notamment la syntaxe et la sémantique, est généralement inexistant. Le travail étant confié à un seul processus.

La mise en oeuvre d'un processus de coopération et d'ordonnancement est difficile et complexe.

- ***Complexes à mettre en oeuvre***

l'excentricité de la langue naturelle et de son ambiguïté : la taille des données nécessaires est infiniment importante [DÉBILI, 82].

Les inconvénients :

Bien que les procédures linguistiques paraissent les mieux appropriées à la résolution des problèmes liés aux traitements du langage, les systèmes de ce type se heurtent à de nombreuses difficultés :

- détermination de la profondeur de l'analyse syntaxique,
- la recherche des sous-unités syntagmatiques qui sont pertinentes pour représenter le document,
- le paraphrasage du vocabulaire d'indexation,
- etc.

Ainsi, la plupart des systèmes d'indexation automatique actuellement commercialisés sont qualifiés de systèmes mixtes : les procédures linguistiques délimitent les unités d'indexation potentielles, tandis que les procédures statistiques effectuent un filtrage et un ajustement d'éventuels poids pour ces unités.

II.5- Thesaurus

Les systèmes d'indexation rencontrent de grandes difficultés dans la localisation des unités de sens chargées de saisir les thèmes abordés dans les documents.

L'établissement de thésaurus constitue donc un des axes de recherche pour résoudre ce type de problème.

II.5.1- Principes généraux

La difficulté pour construire un thésaurus traduit l'importance de la normalisation du vocabulaire d'indexation qui est construit à partir :

- d'une liste de mots fonctionnels,
- d'un dictionnaire des synonymes et des paraphrases,
- d'une hiérarchie des termes.

Celui-ci, support de l'indexation des textes, assure une consistance dans l'établissement des identificateurs de documents.

Quand les textes de la base traitent un sujet récent, le thésaurus n'est pas établi et il devient indispensable de le définir. Or, la construction d'un thésaurus est le résultat de plusieurs années de mise en occurrence des analyses de spécialistes et de documentalistes du domaine.

Les coûts de construction d'un thésaurus (temps et argent) étant largement prohibitifs, les recherches se sont donc tournées vers l'acquisition automatique de ses termes.

Ainsi, nous aborderons la constitution du thésaurus sous ces trois aspects :

1- la sélection des mots le composant ;
2- l'identification de syntagmes pertinents ;
3- la définition d'associations entre ses éléments .

II.5.2- Approches statistiques

Dès 1960, des recherches se sont portées sur les associations statistiques entre les termes. Des « *cartes de termes associés* » (en anglais : term association maps) ont été dressées pour suppléer aux traditionnels termes alphabétiques.

♣ Mesure de la co-occurrence de deux termes :

Deux mots sont considérés comme associés lorsque leur co-occurrence dépasse un seuil qui est fonction de la longueur du texte.

Les différentes mesures de la co-occurrence de deux termes reflètent les approches de certains auteurs dans ce domaine :

- Pour L.B. Doyle, la co-occurrence de deux termes dans un texte reflète la pensée de l'auteur et traduit une association cognitive [DOYLE, 62].
- Pour H.F. Stiles, c'est une fonction de la base documentaire (ensembliste) [STILES,61].

Nous rappelons que la *fréquence* d'un terme est le nombre de ses occurrences dans la totalité du corpus, alors que la *prévalence* est le nombre de textes dans lesquels le terme apparaît.

Soient:

A: la prévalence du terme A ;
B: la prévalence du terme B ;
f : la prévalence conjointe des deux termes A et B (A et B dans le même texte) ;
N: le nombre de texte.

L.B. Doyle mesure la **Co-occurrence (A , B)** = $\frac{f}{A+B-f}$

H.F. Stiles propose le **Facteur d'association (A , B)** = $\log_{10} \frac{(|fN-AB|-\frac{N}{2})^2 N}{AB(N-A)(N-B)}$

La mesure en fonction de N, pour H.F. Stiles, devient plus importante quand le nombre de document augmente : les deux termes A et B ont d'autant plus d'importance qu'ils marquent une différence entre les documents.

L.B. Doyle supprime l'augmentation de N pour rester fidèle à son interprétation cognitive. Sa mesure traduit l'association sémantique entre les termes A et B indépendamment de la collection considérée [WYLLYS, 62].

♣ Acquisition de syntagmes :

L'acquisition d'expressions formées de plusieurs mots est indispensable, car pour certaines expressions (les proverbes, les noms composés, les idiomes, etc.) leur sens global ne peut être déduit du sens de ses composants.

De telles expressions sont souvent absentes des dictionnaires où leur mise au point est difficile pour un lexicographe. L'automatisation de la sélection des expressions (ou mots nouveaux) constitue un projet de recherche [VINESSE, 97]où le modèle formel d'extraction de syntagmes sera le compromis de plusieurs années de recherche et d'expérimentation [BIÉBOW, 97].

II.5.3- Approches linguistiques

Dans la plupart des thésaurus construits automatiquement, les liens entre les mots sont construits par des calculs statistiques. Pourtant, les recherches s'orientent vers l'étude de la sémantique [MANIEZ, 88].

H.F. Stiles a étudié les fréquences conjointes de paires de termes relativement à leurs fréquences individuelles [STILES, 61]. Il a introduit une composante linguistique en lemmatisant les mots des textes analysés. Cette approche permet d'unifier des mots proches en modifiant leur fréquence.

G. Salton, poursuivant la même idée, initie des recherches pour identifier les synonymes, les références des pronoms, etc., dans le but de diminuer les variations du vocabulaire [SALTON,73]. Il préconise l'utilisation de grammaires (grammaire à contexte libre) pour déterminer les limites entre les syntagmes.
La procédure théorique d'analyse utilisée, en anglais « *Sentence Kernelization »*, est divisée en deux passages :

Passage_1: Utilisation conjointe d'un analyseur de structures de syntagmes et d'une grammaire transformationnelle afin de produire l'ensemble des noyaux ou segments de phrases correspondant à des structures syntaxiques simples.

Passage_2 : Identification des noyaux sémantiquement équivalents.

G. Salton identifie une liste de problèmes gênant l'identification des syntagmes :
- des mots isolés sont des synonymes ou des homographes ;
- des constructions syntaxiques différentes sont sémantiquement équivalentes ;
- les références pronominales sont liées à un référent ;
- des relations non exprimées sont compréhensibles par déduction ;
- des constructions se réfèrent à des items non spécifiés ou spécifiés ailleurs.

Dans les années 70, A. Andreewsky et al. ont défini et implanté l'approche des *filtres linguistiques* qui est utilisée dans le système SPIRIT [ANDREEWSKY, 75]. Le système dispose d'une grammaire acquise par apprentissage automatique et d'un dictionnaire [ANDREEWSKY, 96].

Il s'agit de sélectionner des groupes de mots qui correspondent à des filtres particuliers qui sont basés sur l'examen des positions syntaxiques des mots dans la phrase [AGIRRE, 96]. Ce filtrage syntaxique permet de sélectionner les mots composés qu'il est souhaitable d'inclure dans le thésaurus du domaine.

T. Ahlswede étudie de gros corpus de textes pour établir des liens étiquetés entre les mots et pour construire des dictionnaires d'expressions [AHLSWEDE, 88]. Le processus d'acquisition utilise un dictionnaire électronique (11,000 expressions).

L'analyse des expressions permet d'établir des liens entre les éléments du thésaurus et de qualifier la nature de relation établie en tant que relation sémantique (avec quelquefois une intervention humaine) à la place des relations statistiques établies.

II.5.4- Approche connexionniste

La représentation de thésaurus utilisée jusqu'ici s'avère de moins en moins appropriée aux besoins actuels de performance et de pertinence des réponses. Cette représentation est, en effet, mal adaptée à la prise en compte de certains concepts, tels que les relations sémantiques, l'occurrence et la co-occurrence entre items.

L'approche que nous exposons a pour objectif de résoudre ces problèmes en mettant en oeuvre une architecture connexionniste fondée sur les réseaux de neurones [BOUGHANEM, 92], [HAMOU-LHADJ, 94], [DAVALO, 92], [JODOUIN, 94].
.

II.5.4.1- Neurone biologique

Le principe de la perception-action est à la base de plusieurs mécanismes, plus ou moins complexes, du traitement de l'information chez les êtres vivants. Le traitement de l'information se fait grâce au système nerveux central, dont les éléments de base sont les cellules nerveuses, appelées neurones.

Ces neurones possèdent de nombreux points dans leur organisation avec les autres cellules. Ils présentent des caractéristiques qui leur sont propres, à savoir :

- la réception de signaux en provenance des neurones voisins,
- l'intégration de ces signaux,
- la possibilité d'engendrer un influx nerveux,
- le conduire, et
- le transmettre à un autre neurone.

Un neurone est constitué de trois parties :

1- le ***corps cellulaire***, contient le noyau du neurone assurant sa vie ;
2- les ***dendrites***, sont les récepteurs principaux du neurone pour capter les signaux qui lui parviennent ;
3- l'***axone***, sert de moyen de transport des signaux émis par le neurone.

Pour former le système nerveux, les neurones sont connectés les uns aux autres suivant des répartitions spatiales complexes. Les connexions entre deux neurones se font en des endroits appelés ***synapses***. Ainsi, l'information circule sous forme de trains d'impulsions électriques d'amplitude sensiblement constante et de fréquence variable, et les synapses peuvent être excitatrices ou inhibitrices.

Le traitement de l'information se fait par différentes voies, utilisant des procédures souvent parallèles et mettant en jeu simultanément ou successivement des niveaux hiérarchiques différents de reconnaissance et de représentation.

Parmi les mécanismes répertoriés, on peut relever :

- ♣ l'**association** : permet d'instancier une représentation par une autre qui lui est proche.
- ♣ la **classification** : permet de décider de l'appartenance d'un stimulus, éventuellement incomplet ou bruité, à une certaine classe.
  Le but du classement est d'attribuer une valeur à ce stimulus.
- ♣ la **généralisation** : Les mécanismes d'associations et de classifications s'appuient sur des connaissances préalables, acquises explicitement ou implicitement par apprentissage. Dans le cas où les stimuli ne seraient pas explicites, un nouveau mécanisme intervient ayant la capacité de généralisation.

II.5.4.2- Neurone formel

S'inspirant des travaux scientifiques sur les neurones biologiques, McCulloch et Pitts (1943) ont proposé le modèle formel du neurone.
Le neurone formel fait une somme $\{w_1 ,w_2 ,...,w_n\}$ pondérée par des potentiels d'actions $\{e_1 ,e_2 ,...,e_n\}$qui lui parviennent. Chacun de ces potentiels $\{e_i , i=1..n\}$ et de ces poids

{w_i , i=1..n} est une valeur numérique qui représente respectivement l'état du neurone qui l'a émis et l'importance du lien avec ce dernier :

$$S = \sum_i (e_i * \quad w_i)$$

puis le neurone formel s'active suivant la valeur S de cette sommation (*Fig. II.5.4.2.*) :

- **Si** cette somme (S) dépasse un certain seuil (b) **alors** le neurone est activé et transmet une réponse (sous forme de potentiel d'action) dont la valeur (A) est celle de son activation (la fonction d'activation F du neurone).
- **Si** le neurone n'est pas activé **alors** il ne transmet rien.

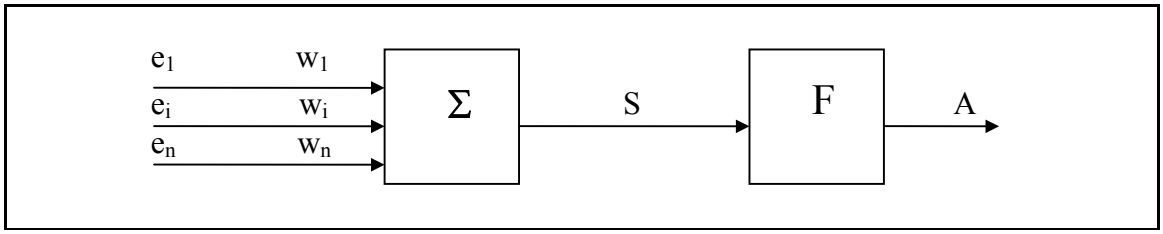


Fig. II.5.4.2. Neurone formel avec fonction d'activation.

Dans le modèle d'origine, la fonction d'activation F est une fonction à seuil. Mais, les chercheurs soucieux de conformité avec le modèle biologique donnent une description continue du neurone.
La fonction F va produire un signal continu qui rend l'importance de l'activation du neurone. Cette fonction est généralement bornée, continue et croissante comme celle des fonctions sigmoïdes :

$$F_1(x) = \frac{1}{1 + e^{-ax}} \quad \text{et} \quad F_2(x) = \frac{1 - e^{-ax}}{1 + e^{-ax}}$$

qui sont respectivement à valeurs dans [0,1] et dans [-1,+1]. Le paramètre (a) est le gain qui mesure la raideur de la sigmoïde.

II.5.4.3- Réseau neuronal

Un réseau neuronal est un ensemble de neurones formels interconnectés et évoluant dans le temps par interactions réciproques. La description statique de la fonction d'activation d'un neurone formel n'est pas suffisante pour caractériser un réseau.

Ainsi, un réseau neuronal se définit par :

- son architecture, qui représente la structure de ses connexions,
- les fonctions d'activation de ses neurones,
- la dynamique de ses connexions.

La structure des connexions dans un réseau neuronal peut aller d'une connectivité totale (les neurones sont connectés entre eux) à une connectivité locale. Les ***poids des connexions*** sont des valeurs qui leur sont attribuées afin de rendre compte de l'influence d'un neurone sur ceux qui sont reliés à sa sortie.

La ***dynamique des états*** correspond à l'évolution des états des différents neurones d'un réseau. Cette dynamique dépend à la fois des fonctions d'activation de chaque neurone, de la structure

et des poids des connexions. L'évolution des poids des connexions correspond à la ***dynamique des connexions***.

Les nombreux modèles de réseaux neuronaux forment deux grandes familles :
- les réseaux non récurrents,
- les réseaux récurrents.

♣ Réseaux non récurrents :

La particularité du réseau non récurrent est qu'il est organisé en couches successives. Le premier modèle opérationnel a été le « Perceptron de Rosenblatt » dans les années 1950.

Dans ce réseau, les connexions ne vont que dans un seul sens de la couche d'entrée vers la couche de sortie. Par ailleurs, il n'y a pas de connexion entre les cellules d'une même couche. Chaque connexion entre les cellules d'association et les cellules de décision est affectée d'un poids. L'apprentissage du « Perceptron » est basé sur la règle de Hebb, c'est-à-dire un apprentissage supervisé qui se fait par correction d'erreur.

Cet apprentissage consiste à présenter au réseau une série d'exemples ou couples (E,S) où E est l'entrée et S est la sortie désirée, puis à minimiser l'erreur entre la sortie désirée S et la sortie effective Y.
L'algorithme est décrit comme suit (*Alg.II.5.3.3.*) :

1. Réseau = ensemble de neurones,

   Pour le neurone j : $\{e_i , i=1..n\}$ et S_j ,

2. Calculer les sorties obtenues: pour le neurone j , $Y_j = \sum_{i \neq j} (e_i * \omega_i)$

3. Calculer les nouveaux poids des connexions :

   pour la connexion entre les neurones i et j ,
   $\omega_{ij}^{(t+1)} = \omega_{ij}^{(t)} + a(S_j - Y_j)$

   ω_{ij} : le poids de la connexion entre i et j
   S_j : la sortie désirée pour j
   Y_j : la sortie obtenue (effective) pour j
   a : coefficient de la vitesse d'apprentissage
   t : le temps
   i,j : neurone i et neurone j

Alg.II.5.3.3. Algorithme d'un réseau non récurrent – le Perceptron.

Pour perfectionner ce modèle, Minsky et Papert (1969) ont ajouté d'autres couches de neurones formels. Ainsi, les sorties des unités d'une couche servent d'entrées aux unités de la couche suivante dans le nouveau modèle.

Inconvénients :

Dans ce type de réseau, la dynamique est implicite, car l'état des unités est donné fonctionnellement par l'état des unités d'entrées. La difficulté dans ce modèle est qu'il n'existe pas de règles ou résultats théoriques permettant de dimensionner le réseau et le calcul d'apprentissage est lent pour atteindre la précision recherchée.

♣ Réseaux récurrents :

Le réseau est dit récurrent s'il existe un circuit dans son graphe orienté de connexion. Il existe deux modèles de réseau récurrent :

- le modèle de Hopfield : réseau entièrement connecté [CHEN, 98b],
- le modèle de Kohonen : réseau partiellement connecté.

➢ Le modèle de Hopfield

Selon Hopfield, le système nerveux recherche des états stables attracteurs dont les états voisins tendent à se rapprocher. Cela permet de corriger les erreurs et de compléter les informations.

Le réseau de Hopfield est constitué par un ensemble de neurones binaires tous reliés les uns aux autres. Chaque neurone j calcule la somme pondérée S_j des excitations de tous les autres neurones et détermine son état :

$$S_j = \sum_{i \neq j} (w_{ij} * A_i) \quad \text{avec } A_j = \begin{cases} 1 & si \quad S_j \rangle 0 \\ 0 & \sin on \end{cases}$$

Si la matrice des poids des connexions est symétrique, la stabilité du réseau est déterminée par le minimum d'une fonction d'énergie, appelé "attracteur", définie par :

$$h = -\frac{1}{2} \sum_i \sum_j (w_{ij} * A_i * A_j)$$

Quand l'état A_i du neurone i est mis à jour, alors : $\Delta h = -\frac{1}{2} \Delta A_i \sum_j (w_{ij} * A_j)$

A chaque itération du calcul, h décroît. Comme h est bornée, il y a donc convergence vers un état attracteur stable : les poids fixes obtenus correspondent à des minima d'énergie.

L'apprentissage dans ce réseau utilise une règle déduite de celle de Hebb, qui garantit une matrice symétrique et les poids des connexions utilisés sont :

$$w_{ij} = \sum_{k=1}^{n} ((2A_i^k - 1) * (2A_j^k - 1))$$

L'apprentissage est effectué une seule fois et à partir de l'échantillon d'apprentissage : l'ajout d'un état stable au réseau est relatif à la modification des w_{ij} par la règle de Hebb.

Inconvénients :

La stabilité du réseau de Hopfield peut être vérifiée par l'étude de la fonction d'énergie du réseau, mais le rappel des exemples appris n'est pas toujours réalisé. De plus le réseau ne permet pas de mémoriser un grand nombre d'exemples : 0.15 fois moins le nombre de neurones.

Selon Hopfield (1984), les résultats obtenus sur le réseau binaire peuvent être étendus au réseau à activité continue.

➢ Le modèle de Kohonen

Le réseau de Kohonen est composé d'une seule couche de neurones, appelée couche d'auto-organisation. Les neurones du réseau sont connectés les uns aux autres par des liens pondérés. Soit w_{ij} la valeur du lien entre les neurones i et j du réseau et aux N composantes du vecteur d'entrée X du réseau.

Chaque neurone de Kohonen reçoit les N composantes du vecteur d'entrée, puis propage le signal vers les autres cellules du réseau. Il existe un mécanisme d'interaction latéral entre les cellules émettrices et celles réceptrices, qui est fonction de la distance entre les neurones : cette fonction est dite "chapeau mexicain".

La discrétisation de la fonction du chapeau mexicain permet de déterminer les poids des connexions internes entre les neurones et la loi d'évolution du neurone est donnée par la fonction suivante :

$$S_i^{(t)} = f(E_i^{(t)} - \sum_{k=-m}^{+m} (l_k * S_{i+j}^{(t-1)})$$

où :

f : une fonction sigmoïde.

$E_i^{(t)}$: la somme pondérée des entrées du neurone i à l'instant t.

l_k : le coefficient de discrétisation de la fonction "chapeau mexicain".

$S_i^{(t)}$: la sortie du neurone i.

t : le temps.

L'interaction latérale entre les neurones regroupe les neurones excités autour du neurone le plus activé par le stimulus. Pour arriver à cette configuration, la règle d'apprentissage classe les vecteurs d'entrée dans des groupes de neurones qui leur sont similaires par ajustements des pondérations entre les entrées et les neurones.

L'algorithme de l'apprentissage en auto-organisation repose sur le processus itératif suivant.

1- ***Sélection de la zone qui répond au vecteur d'entrée***

Le vecteur d'entrée est une image du neurone i dans l'espace des signaux d'entrée. Pour sélectionner un signal qui répond à un type donné de vecteur d'entrée, une comparaison des vecteurs des poids des connexions avec les vecteurs d'entrée afin de pouvoir coupler ceux qui se ressemblent le plus :

$$\left\| X - w_e^t \right\| = \min\left(\left\| X - w_i^t \right\| \right) \;;\; 1 \le i \le N$$

où :

X: le vecteur d'entrée.

w_i^t : le vecteur poids du neurone i.

w_e^t : le vecteur poids du neurone e le plus proche du vecteur d'entrée X.

t : le numéro de l'itération.

N : nombre de neurones du réseau.

2- ***Ajustement des poids entre les neurones sélectionnés***

$$w_i^{t+1} = \begin{cases} w_i^t + a(t)(X - w_i^t) & pour \quad i \in V_e \\ w_i^t & pour \quad i \notin V_e \end{cases}$$

où :

a(t) : coefficient entre [0,1] qui décroît en fonction du temps d'apprentissage t

V_e : l'ensemble de neurones se trouvant au voisinage du neurone e.

t : le numéro de l'itération.

Pour que cette fonction converge, les conditions suivantes doivent être vérifiées :

$$\sum_{t=0}^{\infty} a(t) = \infty \quad \text{et} \quad \sum_{t=0}^{\infty} \left(a(t)\right)^2 < \infty$$

Après un certain nombre d'itérations de la séquence précédente, le réseau converge vers un état organisé.

Le réseau de Kohonen est, en général, utilisé pour les opérations de classification et de regroupement [LELU, 97].

II.5.4.4- Apprentissage dans le réseau

De façon générale, la recherche d'information dans un Système de Recherche d'Information (S.R.I.) consiste à mettre en correspondance une requête et un ensemble de documents, et à signaler si un document répond ou non à cette requête.

La simple comparaison entre les termes d'une requête et ceux des documents ne produit pas toujours des réponses acceptables du point de vue de la pertinence. C'est pourquoi plusieurs améliorations ont été développées au sein des SRI et, parmi les méthodes, le modèle connexionniste et l'apprentissage dans le réseau de neurones.

Le problème consiste alors à trouver les termes à ajouter à la requête pour augmenter la pertinence des réponses en fonction d'associations préétablies entre les différents termes et/ou les différents documents dans le SRI.

Du fait que les améliorations des SRI sont limitées par les aspects suivants :

- l'association entre deux termes est déterminée soit manuellement (par un travail intellectuel), soit par des méthodes statistiques ne garantissant pas la validité de ces relations et l'identification de toutes les relations entre les termes.
- le poids sémantique d'un terme dans un corpus de documents est calculé à la construction du SRI, donc sa modification est relative aux ajouts ou aux suppressions de documents dans la base.
- la réduction d'un texte pour son indexation peut ne pas contenir les mots censés représenter les concepts essentiels contenus dans le document (à l'origine, les documents sont conçus par des personnes qui ne sont pas forcément des linguistes ou des documentalistes).

Pour pallier ces inconvénients et augmenter les performances du SRI, plusieurs fonctionnalités peuvent être rajoutées, parmi les possibilités envisageables :

- créer, modifier ou supprimer automatiquement des relations entre les termes répertoriés dans le thesaurus et les documents. L'appartenance d'un terme ou de plusieurs à un document n'est pas fixe,
- enrichir le descripteur d'un document par l'ajout de nouveaux termes.
- trouver le moyen de transformer la requête de l'utilisateur en tenant compte de la connaissance contenue dans la base pour l'amélioration des résultats de l'interrogation,
- tenir compte des résultats de recherche précédents pour restructurer la base pour la pertinence des documents en réponse à la requête,
- modifier les poids des termes selon les documents pertinents qui sont sélectionnés par ces termes.

Pour résumer, il s'agit de développer un ***SRI dynamique*** : la base d'informations doit en particulier être évolutive permettant la mise en oeuvre d'une recherche associative. Dans cette optique plusieurs travaux ont été effectués par B. Croft [1991] et D. Savoy [1991], basés sur les réseaux inférentiels.

Pour d'autres, ces travaux sont basés sur l'approche connexionniste. L'un des principaux avantages de cette approche est l'apprentissage permanent pour la restructuration des termes en réponse à la recherche de l'information dans la base, donc la mise en oeuvre des fonctionnalités citées ci-dessus.

II.5.4.4.1- Représentation associative de la base d'informations

Pour formaliser la représentation de la base de documents, un modèle connexionniste est un réseau de neurones à trois couches proposé par M. Boughanem et C. Soule-Dupuy (1991-1992) [BOUGHANEM, 92].
Ce modèle est composé par :

- une couche de requêtes, notée Q,
- une couche de documents, notée D,
- une couche de termes d'indexation, notée T.

Les cellules des couches de documents et de requêtes sont toutes deux reliées aux cellules de la couche de termes (*Fig. II.5.4.4.1.*). La couche de requêtes constitue la couche d'entrée du réseau et représente la requête de l'utilisateur. Chaque cellule de cette couche est un terme de la requête.

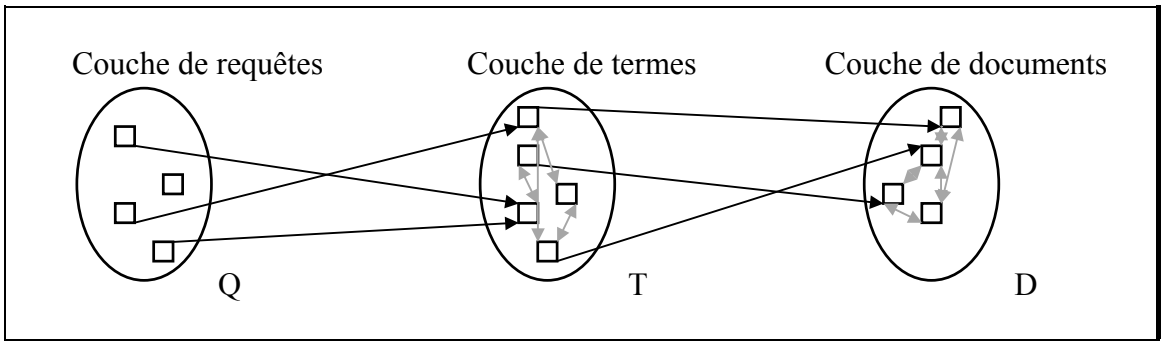


Fig. II.5.4.4.1. Représentation neuronale d'une base de document.

La couche de termes regroupe tous les termes extraits des descripteurs de documents. Elle constitue donc le thésaurus de la base. Une cellule de cette couche représente un ***terme d'indexation***. Ces cellules sont interconnectées par des liens bidirectionnelles valués.

Il faut noter que lors de la phase d'interrogation du système, le thésaurus va subir des changements : ajout de nouveaux mots, modification des liens entre les termes, etc.
Ce thésaurus est qualifié de « ***thésaurus dynamique*** », par opposition au thésaurus classique qui une fois construit ne subit aucune modification.

II.5.4.4.2- Connexion entre deux termes : thésaurus dynamique

Dans la représentation neuronale d'une base de documents, nous distinguons trois types de connexions :

- connexion entre deux termes t_i , t_j

- connexion entre deux documents D_i , D_j
- connexion entre un terme et un document t_i , D_j

Dans ce qui suit, il est question d'examiner la connexion entre deux termes.

Les cellules du réseau de termes sont toutes connectées et le lien entre deux termes t_i et t_j représente un lien d'association. Le but n'est pas d'identifier les différents types de relations qui peuvent exister entre les termes, à savoir la synonymie, la hiérarchie, etc. L'établissement de ces liens reste dans la majorité des cas manuels.

La solution proposée par les auteurs consiste à trouver un moyen automatique qui permettra d'établir des relations entre les mots. Cette opération est réalisée par la prise en compte des documents dans lesquels ces mots apparaissent. Ces relations vont lier des termes qui, lors de la phase d'interrogation de la base, permettent de restituer des documents contenant les mêmes concepts : l'importance d'un lien entre deux termes est proportionnelle à la pondération de ce lien.

La détermination du poids du lien entre deux termes t_i et t_j , à l'état initial du réseau, est basée sur l'hypothèse suivante :

> selon M. Boughanem et C. Soule-Dupuy, *« Le poids d'un lien entre deux termes est augmenté chaque fois que ceux-ci apparaissent dans un même document. Par contre, il se voit diminuer si un seul de ces termes apparaît dans un document. »*
>
> La formule proposée consiste à faire le rapport entre le nombre de documents contenant les deux termes et le nombre de documents contenant chacun de ces termes séparément.
>
> En premier temps la formule s'écrit : $w_{i,j} = \dfrac{card(E_{ti} \cap E_{tj})}{card(E_{ti} \cap E_{tj})}$
>
> où :
>
> E_{ti} : représente l'ensemble de documents contenant le terme t_i
> E_{tj} : représente l'ensemble de documents contenant le terme t_j
> card : représente le cardinal d'un ensemble.

Les différentes valeurs sont regroupées dans une matrice diagonale que l'on appellera matrice d'association W_T : $W_T(i,j) = W_T(j,i) : i, j \in [1..N]$

> W_T représente le lien entre le terme t_i et le terme t_j ,
>
> on peut déduire de la formule que: $0 \le W_T(i,j) = w_{i,j} \le 1$ avec $i, j \in [1..N]$
>
> où N: le nombre de cellules contenues dans la couche T.

Il est clair que cette façon purement statistique de déterminer les relations ne garantit ni l'identification de toutes les relations qui peuvent exister entre les termes, ni la validité des relations trouvées.

C'est la phase d'apprentissage qui va permettre d'affiner ces valeurs : ajouter de nouveaux liens et construire un réseau de termes conforme aux concepts présents dans la base de documents.

II.5.4.4.3- Pondération des liens inter-couches

La base de détermination des poids des liens entre les couches Q-T et les couches D-T est appelée le ***poids sémantique*** d'un terme. Le poids d'un terme t_i va déterminer l'importance de ce terme dans la caractérisation du contenu d'un document ou d'une collection de documents.

La tendance actuelle dans la théorie des réseaux de neurones est d'attribuer aux liens des valeurs proches du résultat que l'on veut obtenir [BRILL, 92].
Dans le cas de l'approche des auteurs [BOUGHANEM, 92], la formule de calcul de poids est développée dans les méthodes statistiques. Cette formule du poids sémantique est utilisée dans le système RECOLTE : $poids(t_i) = (1 - freq(t_i))\sqrt{N} = (1 - \frac{n_i}{N})\sqrt{N}$

Cette formule peut être interprétée de deux façons :

- la fréquence d'apparition absolue, selon que l'on utilise la fréquence d'apparition d'un terme dans un document,
- la fréquence d'apparition relative, selon que l'on considère ce cas dans une collection de documents.

Les valeurs de ces poids seront affinées lors de la phase d'apprentissage, dans le but de refléter l'importance effective de chaque mot dans la caractérisation du contenu de chaque document.

Cette phase d'apprentissage est nécessaire car les poids des termes sont déterminés à partir des formules statistiques qui ne garantissent pas l'exactitude de ces poids.

Par conséquent, l'importance d'un mot dans une collection de documents dépend de son importance par rapport à chaque document.

La relation est la suivante : $q_{ik} = \frac{\sum_{j=1}^{M} p_{ij}}{M}$

où :

q_{ik} : le poids du lien entre la requête Q_k et le terme t_i

p_{ij} : le poids du lien existant entre le terme t_i et le document D_j

M : le nombre de connexions partant de t_i vers D. Il représente également le nombre de documents contenant le terme t_i .

II.5.4.4.4- Apprentissage

La règle d'apprentissage pour modifier les liens entre T et D est celle de Hebb. Elle consiste d'une façon générale :

- à augmenter les poids des liens entre les documents jugés pertinents et les termes activés qui ont permis de les sélectionner,
- à diminuer les poids des liens entre les documents jugés non pertinents et les termes activés qui ont permis de les sélectionner.

La règle de modification de poids est la suivante : $p_{ij}^{(t+1)} = p_{ij}^{(t)} \pm \tau . \Delta p_{ij}^{(t)}$

où

τ : la vitesse d'apprentissage.

$p^{(t)}_{ij}$: le poids de la connexion (t_i, D_j) à l'étape t ,

une étape correspond à une recherche,
$\Delta p^{(t)}_{ij}$: le pas d'apprentissage.

On en déduit que le pouvoir significatif d'un terme par rapport à un document sera soit augmenté soit diminué selon la pertinence du document auquel il est relié. Cette variation va donc impliquer la variation de n_i (nombre de documents contenant le terme t_i).

Dans cette approche, le pas d'apprentissage Δp^{t}_{ij} va représenter le taux de variation de la fonction poids sémantique.

Conclusion :

L'intérêt de cette approche est double. Elle permet d'une part, la construction d'un thésaurus dynamique qui va permettre la création de nouveaux liens de similitude entre les documents ou les associations entre les termes. D'autres part, l'auto-organisation du fonds documentaire au cours de la phase d'apprentissage grâce au mécanisme de recherche associative.

II.6- Conclusion

Nous avons proposé un état de l'art dans le domaine de l'indexation automatique en soulignant dans ce domaine les problèmes rencontrés dans chaque méthode ou approche étudiée.

Certes, les solutions ne sont pas définitives à travers les recherches sur l'indexation automatique des documents, mais les approches semblent résoudre certains problèmes, essentiellement liés aux traitements du langage naturel.

Les problèmes liés aux langues naturelles ne connaissent pas de solutions définitives mais optimales pour certains niveaux du traitement de la langue. Les enjeux sont importants et concernent autant les industriels de la langue que les chercheurs universitaires.

Aussi bien dans le passé qu'à l'avenir, la prolifération des systèmes d'indexation automatiques ou semi-automatiques continuera à se développer. Ces systèmes dédiés à capitaliser les connaissances et l'expérience sur les sources d'information montrent un intérêt prononcé pour la gestion des connaissances et du savoir-faire « de l'entreprise ». Il sera ainsi possible de mettre à la disposition des personnes usagers des techniques « intelligentes » pour leur besoin d'interrogation des bases ou banques documentaires : sources de connaissances.

Un tel service rendu aux usagers constituera un aspect opérationnel permettant d'apporter une aide non négligeable au problème de la conservation, de la gestion et de la pérennité des connaissances de toute entreprise.

CHAPITRE III :

Modèle linguistique pour l'indexation automatique et étude de quelques systèmes

III.1- Introduction

Nous tenterons dans ce chapitre d'exposer un modèle linguistique qui rend compte de la problématique posée en général par les systèmes d'information documentaire et en particulier par l'indexation automatique.

Rappelons qu'indexer un document consiste à lui donner une représentation qui permet de faciliter l'accès à l'information (contenu du document).

Or, dans un état de l'art sur l'indexation automatique documentaire (*cf.* chapitre II), le statut du descripteur ou du mot-clé est souvent corrélé avec des traits et des aspects intrinsèquement linguistiques.

Le fait de penser à une approche linguistique pour l'indexation automatique est une autre conception des systèmes d'information documentaire, car on se rend compte du phénomène de référence à la réalité extra-linguistique (objets du monde réel) :

> *Par ce biais, on arrive à distinguer entre les mots de la langue et ceux du discours.*
>
> *Selon R. Bouché, « les deux approches, langage d'indexation et prise en compte d'une valeur référentielle, apparaissent donc comme complémentaires. »* [BOUCHÉ, 88].

Certaines études linguistiques se limitent aux syntagmes nominaux pour aborder le problème de l'indexation automatique.

> Le syntagme nominal est en effet *« l'unité minimale du discours qui permet de désigner un objet»* [LE GUERN, 89].

Dans la mesure où la langue se définit par l'ensemble des règles universellement présentes dans la communauté « linguiste », il est concevable que les mécanismes qui la caractérisent aient été recherchés au niveau des combinaisons et substitutions élémentaires desquelles toute parole est possible. Ainsi, il semble convaincant d'utiliser les syntagmes nominaux comme descripteurs, ils sont les thèmes explicites des composantes du texte à indexer. En termes empruntés à la logique, on peut dire que la normalité locale qui contrôle la production d'un type de thème donné (ou discours) concerne non seulement la nature des prédicats attribués à un sujet, mais aussi les transformations que ces prédicats subissent au fil du discours [PÊCHEUX, 69].

Il résulte de ce processus discursif, la supposition de deux ordres de recherche : - l'étude des variations spécifiques liées aux processus de production sur le fond invariant de la langue (sémantique, rhétorique et pragmatique), et – l'étude de la liaison entre les conditions de production du discours. C'est ce dernier aspect que nous allons tenter d'éclairer, à travers un examen critique sur la nature du concept « descripteur » dans un processus d'indexation et de recherche d'information dans les bases de données textuelles.

III.2- Théorie linguistique

III.2.1- Traitement automatique des langues

L'objectif d'un traitement automatique des langues naturelles est la conception de programmes (ou logiciels) capables de traiter de façon automatique des données linguistiques.

Qui dit *traitement* dit manipulation d'un objet d'entrée, aboutissant à la modification de cet objet en objet de sortie. Le traitement que nous visons ici agit sur un texte soit pour le transformer, soit pour en extraire de l'information.

Or, pour pouvoir traiter un objet, il faut connaître les principes de sa constitution interne permettant de le décrire de façon opératoire :

> *Le texte est un ensemble de formes et de sens régi par des règles explicites. Ces règles sont les règles de la langue de ce texte.*

Ainsi, le traitement automatique se trouve entraîné à décrire en tout ou partie la langue qui constitue le support du texte à traiter.

Les réalisations concernant l'écrit ont trait principalement aux traitements de textes, aux systèmes d'extraction d'informations et aux systèmes de traduction. L'extraction d'informations consiste à alimenter des systèmes informatiques, en particulier des bases de données, à partir d'informations contenues dans des documents [BOUCHÉ, 91].

Ces systèmes peuvent être très frustres et se contenter de repérer des mots, et d'autres, des groupes de mots ou des paraphrases qui servent à indexer ces documents.

III.2.2- Enjeux théoriques

Le linguiste sait comment fonctionne la langue, l'informaticien sait comment marche l'ordinateur : dès lors, apprendre à l'ordinateur les règles de la langue pour lui faire traiter des données linguistiques revient, d'une part, à exprimer les règles de la langue qu'un linguiste a inconsciemment intériorisé, et d'autre part, à assimiler la langue naturelle à un langage artificiel (programmable et calculable) qui sera compréhensible par l'ordinateur.

Cette situation très particulière conduit à affirmer des conduites diamétralement opposées [FUCHS, 93] :

- L'informaticien, s'appuyant sur sa compétence, se croit « naturellement » capable de décrire la langue, qu'il tend à aborder de façon réductrice à l'image d'un langage formel.
- Le linguiste de son côté évite de tomber dans le travers d'une attitude de fusion subjective avec l'objet. Il a une conscience aiguë de la complexité et de

l'hétérogénéité des facteurs constitutifs de la langue. Il est enfin habitué à rencontrer des phénomènes linguistiques complexes dont la formalisation est au dessus de toute mise en équations.

En retour de cette situation, une insatisfaction de l'informaticien, constatant que les règles de fonctionnement de la langue sont loin d'être complètes et opératoires.

> *Chaque école linguistique avance sa propre conception de la langue, choisit son terrain privilégié de faits à décrire et opte pour une démarche théorique particulière.*
> *Par conséquent, l'informaticien se trouve confronté à des descriptions rivales et incomparables, parce que leurs concrétisations sont difficilement généralisables.*

III.2.3- Couverture de la langue

Toute langue naturelle possède en elle-même un potentiel d'expression illimité, ce qui constitue un obstacle à son traitement automatique [LALLICH, 90] :

> *« un dictionnaire ne sera jamais complet et plus grave encore, aucune modélisation aussi générale soit-elle ne couvrira tous les phénomènes linguistiques ».*

La langue naturelle possède outre son potentiel illimité, deux autres caractéristiques qui constituent des obstacles à son traitement :

- son ***excentricité***, qui se traduit par le fait que le traitement de la majeure partie des phénomènes de la langue avec des données de taille modeste. Par contre, pour la partie restante, la taille des données nécessaires est infiniment plus importante : 100 mots pour reconnaître 60% des mots du texte, 1000 pour 85%, 4000 pour 97%, 50000 pour 99%.
  Aussi, la syntaxe relève-t-elle du même phénomène [CAVAZZA, 92].
- son ***ambiguïté***, qui est caractéristique de la langue elle-même. Cette ambiguïté est source d'une explosion combinatoire si on ne limite pas le nombre de solutions à chaque étape de l'analyse [CARMONA, 98]. Le modèle linguistique concourt à lever cette ambiguïté de plusieurs manières.

III.2.4- Un modèle calculable

L'étude théorique d'un modèle linguistique sur lequel se fonde notre approche pour l'indexation automatique de documents doit posséder la propriété d'être calculable [GROSS, 96].

Cette propriété conduit à la mise en équation des principales caractéristiques de la langue, donc à une description algorithmique cohérente.
C'est la raison pour laquelle, le modèle linguistique doit avoir recours à un modèle formel implémentable, résultant de l'interaction [JACQUEMIN, 2000] entre les aspects linguistiques et algorithmiques (représentations calculables aux descriptions [HABERT, 91-97]).

III.3- Une approche linguistique à l'indexation automatique

Notre hypothèse de travail est que l'analyse d'un énoncé en langage naturel ne peut s'opérer sans faire appel à des fondements théoriques. La linguistique est la science la plus apte à proposer ses modèles pour des données de nature textuelle :

" *... le recours à la linguistique est le seul guide sûr dans le passage des formes de surfaces au codage recherché : seules les procédures linguistiques introduisent dans la démarche une rigueur suffisante pour* ***catégoriser****,* ***regrouper*** *et* ***interpréter***." [ROUAULT, 88].

Étant donné que la construction des énoncés est importante pour notre étude, l'analyse morphosyntaxique dont il sera question devrait permettre de repérer des expressions particulières (syntagmes nominaux) dans un texte écrit et d'en faire usage dans le contexte de l'indexation documentaire et de la recherche d'information.

III.3.1- Stratégie du groupe SYDO-Lyon

L'équipe SYDO-Lyon s'est beaucoup intéressée à l'analyse morpho-syntaxique de la langue française [BOUCHÉ, 90] et à la représentation de textes écrits en vue de définir un modèle général des systèmes d'informations [LAROUK, 93a-b] : bibliothèque, centre de documentation, base de données bibliographiques, base de données textuelles, etc.

Les chercheurs de l'équipe ont conçu une grammaire de reconnaissance des syntagmes nominaux (SN). Cette grammaire s'inspire des grammaires génératives dans la mesure de la reconnaissance seulement.

La solution adoptée est la définition de règles de réécritures des syntagmes de la langue française, qui est à la base de la construction de l'analyseur. Ces règles permettent d'extraire les SN et de construire une base d'information structurée autour du SN.

Une grille de catégories morphologiques est construite pour les règles de réécritures. Le lexique de l'analyseur fournit les éléments d'informations nécessaires dans les ordres morphologique et syntaxique.

III.3.2- Syntagme nominal et descripteur

Pour arriver à designer ce que nous entendons par un descripteur, nous sommes partis de la notion de « ***terme*** ».
Au niveau linguistique, un terme est l'unité qui sert à désigner un concept appartenant à une discipline particulière.

Au niveau de l'indexation documentaire, il est l'unité qui sert à l'indexation des documents dans un système d'information documentaire, aussi appelé ***"descripteur"***.

> Selon W. Mustafa Elhadi, le terme ne peut être autre chose qu'un « terme préférentiel » que choisira le documentaliste parmi tant d'autres qui se trouvent être autant de candidats descripteurs [MUSTAFA, 89].

Le « terme descripteur » peut être, dans une certaine perspective, le synonyme de terme :

> " *Descripteur et terme renvoient à la même réalité et sont donc une relation de synonymie référentielle. Cette synonymie est due à une variation de facette comme le dit M. Le Guern, c'est-à-dire selon le point de vue sous lequel on considère le concept désigné par les deux termes; en traduction, ce concept figure sous l'étiquette* ***terme****, en documentation, en revanche, il figure sous l'étiquette* ***descripteur*** *et son rôle est la représentation du monde.*" [MUSTAFA, 89].

Aussi, il convient de distinguer les descripteurs aux mots du lexique :

M. Le Guern établit une comparaison entre le mot de la langue et le descripteur et met en évidence la différence entre la synonymie lexicale et la synonymie référentielle :

" *La prise en compte de la fonction référentielle des descripteurs permet de poser en d'autres termes la question de synonymie : deux descripteurs sont synonymes s'ils ont la même référence; il ne s'agit donc pas, dans une perspective documentaire, de synonymie référentielle, alors que la seule prise des signifiés linguistiques conduirait plutôt à y voir une certaine antonymie.*" [LE GUERN, 89]

III.3.3- Description du syntagme nominal

Selon M. Le Guern, le syntagme nominal est :

" *L'unité minimale de discours qui a la possibilité de signifier un objet (...) {MAISON}, le mot du lexique, ne signifie aucune maison que ce soit, alors qu'il suffit que le discours construise le syntagme {UNE MAISON} pour que soit désigné un objet concret. La fermeture du prédicat par le quantificateur {UNE} le transforme en terme.*" [LE GUERN, 89].

« Maison » en tant que mot du lexique, est considéré par l'auteur comme prédicat libre qui ne suppose aucun univers déterminé. Le lexique concerne les mots indépendamment des choses. Le passage du prédicat libre au prédicat lié est une opération qui consiste à placer le mot du lexique dans un univers de discours.

M. Le Guern poursuit la description du SN en disant :

" *Je prends dans le lexique le prédicat libre {MAISON} et je le place dans mon univers de discours ; de ce fait, il n'est plus libre, il devient un prédicat lié, et je suis passé dans une* ***logique extensionnelle****. Transformer un prédicat libre en prédicat lié, c'est lui associer une classe d'objets pris dans un univers déterminé*.", [LE GUERN, 89].

Ainsi, la description d'un SN réside dans son statut référentiel, qui est le segment de réalité qui lui est associé. L'autonomie grammaticale et la fermeture au sens de la logique constituent pour le SN d'excellents points d'accès à l'information [LE GUERN, 82].

III.4- Modèle théorique

L'analyse morpho-syntaxique porte sur des textes. Les textes sont composés d'un certain nombre d'unités linguistiques, dont les plus remarquables sont ceux qui réfèrent à une réalité des objets du monde réel (extra-linguistique). Ces unités remarquables sont des termes ou syntagmes nominaux.

Les mots de la langue ne signifient que des propriétés et jamais des entités ou objet du monde réel. Ils signifient des attributs et non des substances, tant qu'ils ne sont pas mis en oeuvre dans un univers de discours.
Cependant, l'analyse morpho-syntaxique a besoin de caractériser les unités linguistiques d'un texte en connaissant un certain nombre d'informations qui se rapportent à eux.

III.4.1- Logique intensionnelle et logique extensionnelle

Sur le plan logique, le cerveau humain a la possibilité de fonctionner selon deux systèmes différents : la logique intensionnelle et la logique extensionnelle.

La logique intensionnelle a la particularité d'être une logique sans univers de référence, c'est le cas du fonctionnement du lexique d'une langue naturelle. Dès lors, le lexique devient un ensemble d'éléments qui ne sont pas en relation avec des objets (*Fig.III.4.1.*).

Un élément de cet ensemble désigne un prédicat libre. Ce prédicat désigne une propriété et non un objet du monde réel :

> " *Le prédicat libre ne désigne pas une substance , mais une propriété (...)*
> *Au niveau du lexique, on a quelque chose de l'ordre du type. Cette notion de type est étayée avec une autre terminologie qui se retrouve dans le système de Peirce. Chaque nouvelle occurrence d'un lexème donné constitue pour Peirce un signe distinct, un sinsigne. Tous ces sinsignes sont eux-mêmes distincts du lexème en langue. Le lexème en tant qu'il appartient à la langue est un légisigne : ce n'est pas lui qu'on retrouve dans les emplois du discours.*
> *Ce légisigne, le lexème en langue, est le premier interprétant des sinsignes que sont les occurrences du lexème.*" [METZGER, 88]

Type de logique	Univers	Éléments centraux	Type d'opération	Exemples
LOGIQUE intensionnelle	Lexique	Prédicats libres simples ou complexes (des propriétés)	d'un univers	Maison, Village, etc.
LOGIQUE extensionnelle	Univers du discours	Termes ou prédicats liés (classes d'objets)	Quantification (opérateur)	La (maison), Le (village), etc. opérateur: le/ la/ etc. opérande: maison/ village/…

Fig.III.4.1. Logiques intensionnelle et extensionnelle.

Ainsi, à cette étape, on se retrouve au niveau de la logique extensionnelle qui traduit la prise en compte d'un univers du discours. Les éléments de cet univers sont des « termes » ou « syntagmes nominaux ». Ce sont tous des unités minimales du discours susceptibles de pointer sur les objets de l'univers en question des prédicats liés. Un prédicat lié traduit la prise en compte d'un objet dans une classe d'objets.

Le passage du niveau prédicat libre à celui de prédicat lié, se fait par une opération de quantification. Tout syntagme nominal se décompose en deux éléments dont le premier est un prédéterminant jouant le rôle de quantificateur (opérateur) et le second un opérande [DUPONT, 83].

Ainsi, les contours du lexique sont clairement définis et les propriétés exprimées dans le lexique ne sont mises en relation avec des objets d'un univers du discours que dans la mesure où une structure syntaxique intervient.

III.4.2- Le modèle linguistique

L'objectif d'un tel modèle est de permettre l'identification des syntagmes nominaux, tout en mettant en évidence la transition entre les mots du lexique (prédicats libres) et les syntagmes nominaux (prédicats liés) qui pointent sur des objets de la réalité extra-linguistique. Cette transition s'effectue à travers la structure syntaxique qui reconnaît les SN (*Fig.III.4.2.*).

Selon R. Bouché [BOUCHÉ, 89], le modèle conçu a comme objectifs :

- permettre l'identification des SN,
- déterminer la structure de ces syntagmes en mettant en évidence les relations entre ses constituants. Ceci permet le stockage d'une représentation du SN, donc facilite la recherche de l'information,
- mettre en oeuvre le mécanisme de passage de la logique intensionnelle (les mots qui appartiennent au lexique de la langue) à la logique extensionnelle, en arrivant à l'unité à valeur référentielle (le SN).

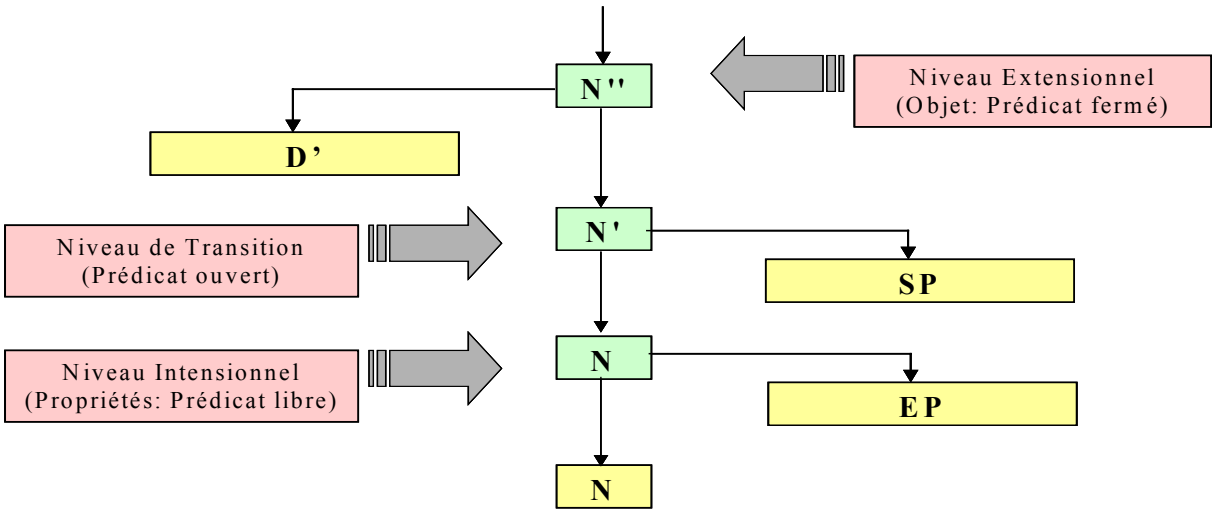


Fig.III.4.2. Grammaire du syntagme nominal.

La grammaire de reconnaissance du syntagme nominal s'articule autour de trois niveaux :

- Niveau **intensionnel** : les propriétés de la langue.
- Niveau **intermédiaire** : la transition de l'intensionnel vers l'extensionnel.
- Niveau **extensionnel** : les objets du monde.

Nous distinguons :

- **N :** niveau purement intensionnel. Les unités considérées sont des prédicats libres simples (les propriétés du nom) ou complexes (les propriétés du nom sont modifiées par des éléments adjectivaux ou des expansions prépositionnelles).
  Exemples : maison[N], *fenêtre*[N], *pomme de terre*[N], *etc.*

- **N' :** niveau de transition entre l'intensionnel et l'extensionnel. Il s'agit de la prise en compte de l'univers du discours considéré. L'intervention de syntagmes

prépositionnels, qui introduit des éléments dont on peut présupposer l'existence, définit une classe d'objets de la réalité extra-linguistique.
Exemples : fenêtre de la maison, maison de Jean, etc.

Selon M. Le Guern et R. Bouché, N' est un prédicat libre mais lié à une classe d'objet (celle de N'').

- **N'' :** au niveau extensionnel, l'opération de fermeture au moyen d'un quantificateur sélectionne un élément précis dans la classe N. Donc une référence à un objet de la réalité extra-linguistique.
Exemples : le placard de cuisine$_{[N'']}$, *le virus du SIDA*$_{[N''][N'']}$ *, etc.*

Le processus d'indexation et de recherche sera évidemment centré sur la mise en oeuvre du mécanisme de référence à la réalité extra-linguistique. Une analyse identique à celle du document permet d'identifier dans la question exprimée en langue naturelle les syntagmes, c'est-à-dire les composantes de syntagmes et les relations qui les lient.

III.4.3- Processus de l'analyse morpho-syntaxique

La transcription du modèle linguistique est la réalisation d'un analyseur morpho-syntaxique du français. Ce qui paraît simple à décrire est loin de l'être en réalité, car l'exécution de ce travail nous oblige à exposer des aspects distincts de la langue qui se complètent entre eux à travers notre démarche :

- le lexique,
- le pré-traitement morpho-syntaxique,
- le traitement morphologique,
- l'analyse syntaxique,
- l'extraction automatique de SN.

Le contexte de coopération entre ces modules est marqué par la conception d'outils informatiques nécessaires au traitement du langage naturel.
Par conséquent, l'analyseur est destiné à opérer avec une grammaire. La grammaire qui sera employée doit être celle d'un système orthographique où les régularités de l'écrit sont formalisées.

L'analyse morphologique doit fournir les données nécessaires aux composants ultérieurs, à savoir : *le module d'analyse syntaxique* et *le module d'indexation automatique*.

L'analyse morpho-syntaxique se déroule sur deux niveaux, l'un préconise une consultation directe du lexique, l'autre, un prétraitement morpho-syntaxique.

III.4.3.1- Le lexique

La création d'un lexique a pour but de faire fonctionner l'analyseur de reconnaissance des SN. Le lexique devrait être fondé sur le principe des "profils" : lexical, syntaxique et flexionnel.

Afin de faciliter le repérage lors de l'analyse des mots composés et des idiomes, nous avons pensé à une méthode souple qui consiste à les intégrer dans un nouveau lexique [FARWELL, 93], [FAURE, 98].

III.4.3.2- La morphologie

La morphologie se divise en trois grandes étapes :

1- une phase de prétraitement morpho-graphique, qui opère sur le texte initial et le régularise. Ce texte prétraité est alors soumis à la deuxième phase,
2- une phase de prétraitement morpho-syntaxique. Celle-ci considère chaque forme du texte et lui associe au moins une entrée lexicale et une catégorie syntaxique. Cette analyse traite chaque forme isolée de son contexte et produit de nombreuses solutions. C'est d'ailleurs la raison d'être de la phase suivante,
3- une phase d'élimination des solutions parasites.

Ces régularisations faciliteront beaucoup la segmentation des phrases en propositions.

III.4.3.2.1- Les prétraitements morpho-graphiques

Ils consistent à substituer une chaîne de caractères par une autre. En règle générale, cette substitution peut-être automatique dans les cas simples ou assistée dans les cas contraires : le texte initial est régularisé pour la suite des traitements.

Parmi ces traitements :

- le codage (ASCII-standard, ASCII-étendu, etc.),
- traitement des majuscules,
- traitement des ponctuations,
- traitement des dates, chiffres, et acronymes,
- etc.

III.4.3.2.2- Les prétraitements morpho-syntaxiques

Ces traitements consistent à la fois à régulariser la surface du texte tout en amorçant l'analyse syntaxique. En effet, la langue française est riche d'amalgames, c'est-à-dire de formes qui résultent du regroupement de deux ou plusieurs formes primaires, chacune ayant un rôle syntaxique propre.

La solution adoptée consiste à se donner un nombre très restreint de catégories syntaxiques, chacune ayant un comportement distributionnel bien défini {V,F,Y,D,P,Q,C,W,T}, voir paragraphe au chapitre III.4.3.4.1-.
Le prétraitement de nature morpho-syntaxique précède brièvement l'analyse morphologique dans le but de détecter, dans les séquences de formes, une propriété syntaxique quelconque. Par exemple l'occurrence de la forme {/ce/ + relatif} est de nature pronominale et non prédéterminative [DE BRITO, 91].

III.4.3.3- Le traitement morphologique

Une analyse morphologique est un automate qui traite isolément chaque forme d'un texte en lui associant des traits informationnels (ou propriétés).

L'analyse morphologique combine deux fonctions :

1- une ***fonction classificatoire***, de nature lexicale, qui consiste à attribuer à chaque forme du texte une catégorie syntaxique,
2- une ***fonction calculatoire*** ou ***flexionnelle***, qui consiste, à partir d'une forme donnée, à calculer sa base pour accéder au lexique.

Une solution consiste à calculer comment une forme provient d'une base, à condition qu'il existe un dictionnaire qui donne pour chaque base, le modèle de son comportement flexionnel.

Le modèle de découpage d'une forme en {Base + Flexions} n'est pas un calcul purement combinatoire, il obéit à des règles d'ordre linguistique défini par : A. Berrendonner en 1983 et complété par M. Le Guern et J-P Metzger en 1988.

Le but du traitement morphologique consiste à dégager un maximum de solutions en surface fondées sur un certain nombre de considérations d'ordre syntaxique [FAY-VARNIER, 91] : la facilité à pouvoir modifier la grammaire du SN sans devoir modifier le traitement morphologique.

III.4.3.4- L'analyse syntaxique

La syntaxe contient, comme la morphologie, une phase de prétraitement suivie d'une phase d'analyse. Les prétraitements consistent essentiellement à regrouper les morphèmes discontinus (les verbes) et à segmenter la phrase en propositions.

L'analyse syntaxique en constituants permet de segmenter la phrase en propositions afin de limiter la longueur de la chaîne à analyser, et la grammaire aux phénomènes internes liés à la proposition.

La chaîne à analyser est une séquence de couples (entrée lexicale, catégorie), correspondant à une proposition. Les seules ambiguïtés qui demeurent sont internes à une catégorie. Le résultat de l'analyse de la chaîne est une arborescence de structures syntaxiques attestées par la langue.

III.4.3.4.1- Grammaire de la proposition

La grammaire s'exprime de façon classique au moyen de symboles et de règles. Les symboles terminaux sont des catégories morphologiques. Les règles peuvent faire intervenir, outre les catégories morphologiques, les variables associées à ces catégories pour compléter les conditions d'application de la règle.

La grammaire du syntagme nominal (SN) a servi de support à plusieurs travaux de recherche au sein du groupe SYDO-Lyon. Le côté gauche de chaque règle est séparé de son côté droit par une flèche ($\rightarrow$), la concaténation est présentée par le symbole '+'.

- $\mathbf{V_T}$: Vocabulaire Terminal de SN

{

Symbole	Catégorie
F-NOM	les noms
F-NOM-PRP	les noms propres
F-NOM-PRO	les noms pronoms
F-NAN	selon le contexte, nom ou adjectif
F-ADJ	les adjectifs
D	les prédéterminants

D-DEF	les prédéterminants définis
D-NUM	les prédéterminants numéraux , cardinaux et assimilées
D-IND	les autres prédéterminants
W-QUA	les adverbes de quantité
W-AAJ	les adverbes d'intensité (modificateurs d'adjectif)
P	les prépositions
P-DE	la préposition /de/
CI, LA	les mots /ci/ et /là/

}.

- $\mathbf{V_N}$: Vocabulaire Non-terminal de SN {N'', N', N, A', A, D', Ep, Sp}

Symbole	Catégorie
N'', N', N, A', A, D', PHR N''_c, A_c, P_c, PHR_c, W_c, SP_c, EP_c,	Est l'axiome. N'' représente la catégorie des syntagmes nominaux. N'' domine N' qui domine N.
EP	est l'expansion prépositionnelle
SP	est le syntagme prépositionnel

III.4.3.4.2- Les règles de la grammaire SN

Description	N° règle	Règle
syntagmes nominaux :	1	N''→ D'+N+F-PRP
	2	N''→ D'+N'
	3	N''→ NOM-PRO
	4	N''→ NOM-PRP
expressions nominales :	5	N'→ N+ SP^n
	6	N'→ N+A'
	7	N'→ N+CI
	8	N'→ N+LA
	9	N'→ N
expressions prédéterminatives :	10	D'→ D-DEF+D-NUM
	11	D'→ P-DE+D-DEF
	12	D'→ W-QUA+P-DE+D-DEF
	13	D'→ W-QUA+P-DE
	14	D'→ D
centres adjectivaux :	15	A'→ W-AAJ+A
	16	A'→ A+EP
	17	A'→ F-ADJ,REL
	18	A'→ A

centres nominaux :	19	$N \rightarrow N+EP$
	20	$N \rightarrow N+A(QUA)$
	21	$N \rightarrow A(QUA) + N$
nominaux :	22	$N \rightarrow$ F-NOM
	23	$N \rightarrow$ F-NAN
	24	$A \rightarrow$ F-NAN,(QUA)
	25	$A \rightarrow$ F-ADJ,(QUA)
syntagme prépositionnel :	26	$SP \rightarrow P+ N''$
expansion prépositionnelle :	27	$EP \rightarrow P+N'$
coordination des adjectifs : - propriétés communes (QUA1)=(QUA2)	28	$A(QUA) \rightarrow A(QUA1)+C+A(QUA2)$
- propriétés différentes (QUA1)≠(QUA2)	29	$A_C(QUA) \rightarrow A(QUA1)+C+A(QUA2)$
coordination de catégories : - syntagme nominal	a	$N''_C \rightarrow N'' + C + N''$
- syntagme verbal	b	$SV_C \rightarrow SV + C + SV$
- verbe	c	$V_C \rightarrow V + C + V$
- adjectif	d	$A_C \rightarrow A + C + A$
- adverbe	e	$W_C \rightarrow W + C + W$
- syntagme préposionnel	f	$SP_C \rightarrow SP + C + SP$
- préposition	g	$P_C \rightarrow P + C + P$
- phrase	h	$PHR_C \rightarrow PHR + C + PHR$

Tab.III.4.3.4.2. Grammaire de réécriture du SN.

Ce modèle linguistique (*Tab.III.4.3.4.2.*) se complète par une représentation sous forme d'une grammaire formelle et son implémentation.

Exemples :

N°règle	Illustration	Règle
1	le + président + Chirac	$N'' \rightarrow D'+N+$F-PRP
2	le + président	$N'' \rightarrow D'+N'$
3	lui	$N'' \rightarrow$ NOM-PRO
4	Chirac	$N'' \rightarrow$ NOM-PRP
5	(la) vente de produit + à la cantine + de l'université	$N' \rightarrow N+ SP^n$, n : entier
6	étudiant + assidu aux cours	$N' \rightarrow N+A'$
7	chien-ci	$N' \rightarrow N+CI$
8	chien-là	$N' \rightarrow N+LA$

9	ministre	N'→ N
10	les + trois (candidats)	D'→ D-DEF+D-NUM
11	de + ces (élections)	D'→ P-DE+D-DEF
12	beaucoup + de + leur (temps)	D'→ W-QUA+P-DE+D-DEF
13	peu + de (résultat)	D'→ W-QUA+P-DE
14	le	D'→ D
15	particulièrement + fidèle	A'→ W-AAJ+A
16	rouge + de colère	A'→ A+EP
17	rectoral	A'→ F-ADJ,REL
18	assidu + aux cours	A'→ A + SP
19	chef + de gare	N→ N+EP
20	drapeau + blanc	N→ N+A(QUA)
21	grand + sportif	N→ A(QUA) + N
22	ville	N→ F-NOM
23	fenêtre	N→ F-NAN
24	joli	A→ F-NAN,(QUA)
25	impartial	A→ F-ADJ,(QUA)
26	(le directeur) de + la gare	SP→ P+ N''
27	(le chef) de + gare	EP→ P+N'
28	(le drapeau) blanc + et + rouge	A_c(QUA)→ A(QUA1)+et+A(QUA2) c: propriétés communes
29	(un plat) chaud + et + parfumé	A_d(QUA)→ A(QUA1)+et+A(QUA2) d: propriétés différentes

III.4.4- Avantages d'une telle représentation

Le fait d'indexer un document par l'ensemble des syntagmes nominaux qu'il contient, représente de nombreux avantages par rapport aux systèmes documentaires classiques. En effet, cette représentation est riche en information et est représentative du contenu.

La non-détection de synonymies entre les syntagmes nominaux est un facteur de silence. La non-prise en compte des relations d'hyperonymie/hyponymie va dans le même sens.

Pour lutter contre le silence dans ce tel mode d'indexation, il est nécessaire de rendre compte des relations de synonymie et d'hyperonymie/hyponymie entre les syntagmes. Dans certains cas, ces relations sont perceptibles à travers l'analyse morpho-syntaxique et dans d'autres cas, il est nécessaire d'intégrer dans le travail un thésaurus en permutant certains termes du syntagme, puis de relancer à nouveau la recherche.

III.4- Description de quelques systèmes d'indexation automatique

Aujourd'hui, les industries de la langue jouent un rôle essentiel dans l'industrie du logiciel. Les logiciels se sont développés selon les problèmes théoriques auxquels étaient confrontés des projets sur le traitement de la langue naturelle et, tout particulièrement, sur le traitement de l'information documentaire.

Dans ce qui suit, on décrit quelques systèmes documentaires ou des prototypes qui font appel à certaines méthodes ou approches (linguistique, statistique, apprentissage, etc.) qui s'intègrent dans le cadre de l'indexation automatique pour le repérage des éléments informatifs du texte.

III.4.1- SPIRIT et MICRO-MIND

Introduction :

L'intelligence artificielle tente de simuler sur ordinateur les comportements de l'homme et les activités liées à sa compréhension. C'est pourquoi l'étude de la langue naturelle joue un rôle si important.

Le cas du système SPIRIT (Système Syntaxique-Probabiliste d'Indexation et de Recherche d'Informations Textuelles), que nous décrivons ci-dessous, est assez représentatif d'une synthèse entre méthodes statistiques et procédures d'analyse syntaxique ou sémantico-lexicale [ANDREEWSKY, 96].

A partir des années 80, avec l'apparition de la bureautique, il a été possible d'envisager la saisie sur support informatique de toute la production littéraire, scientifique ou autre, et de traiter alors le texte intégral par des moyens entièrement automatisés.

Conçus et réalisés au cours des années 1967-75, le système SPIRIT et plus tard MICRO-MIND (1985) répondaient à ce besoin, avec une certaine anticipation sur le développement des modes de communication entre l'homme et l'ordinateur.

Élaboré en vue de l'interrogation en langage naturelle de base de données textuelles, SPIRIT permet grâce à des procédures linguistiques et statistiques entièrement automatisées, d'obtenir pour n'importe quelle base, les documents ayant les contenus les "plus" proches de celui d'une question formulée en langage naturel :

> *« on donne à cette opération le nom d'indexation automatique et interrogation en langage naturel du texte intégral. »* [ANDREEWSKY, 96]

Les premières études qui ont servi de fondement théorique aux systèmes SPIRIT et MICRO-MIND datent des années 1967-73. Elles constituent une réponse aux questions suivantes :

- Peut-on introduire l'utilisation du langage naturel dans l'art documentaire ?
- Comment définir et déterminer automatiquement les termes (ou traits) discriminants d'un document faisant partie d'une base de textes ?
- Quel modèle mathématique faut-il choisir ?
- A ce modèle faut-il adjoindre un modèle linguistique ? et lequel ?
- Peut-on, à partir d'une question formulée dans un langage naturel, obtenir en priorité les documents qui répondent à cette question ?

Un premier modèle mathématique a été développé à partir des modèles Bayesiens qui a permis d'automatiser la sélection et la pondération des termes des textes d'une base.
Des procédures linguistiques étaient perçues comme indispensables pour la cohérence de la sélection effectuée par les algorithmes statistiques.

III.4.1.1- Interrogation en langage naturel

Le problème posé était le suivant :

> *Disposant d'une base de données (textes) sur un support informatique, essayer à partir d'une question posée en langage naturel, d'obtenir par des moyens entièrement automatisés, les textes dont le contenu est "le plus proche" ou "s'apparente le plus" à celui de la question posée.*

La première difficulté de cette approche est celle de la définition de la proximité sémantique de deux textes, c'est-à-dire le texte de la question et celui du document. Comment formaliser cette proximité et les conditions d'automatisation du processus de comparaison sémantique ?

Pour ce faire, la formalisation de certains mots qui peuvent être considérés comme ***plus signifiants*** est telle qu'en leur attribuant une ***pondération appropriée,*** elle permet d'évaluer les proximités sémantiques entre une question et les textes d'une base donnée.

La qualité de cette proximité peut être améliorée si l'on parvient à tenir compte des relations de dépendances entre mots signifiants, c'est-à-dire des propriétés contextuelles.

III.4.1.2- Mots signifiants, pondération et relations de dépendances

L'hypothèse la plus simple qui s'est révélée, par la suite, la plus productive est la suivante :

> Seuls les substantifs, adjectifs, verbes et adverbes (catégories appelées ***mots sémantiques***) pouvaient être des mots plus ou moins signifiants, et que les articles, les verbes auxiliaires, les conjonctions, les prépositions et les pronoms (catégories appelés ***mots fonctionnels***) étaient toujours des mots non signifiants.

Ensuite, le caractère des mots sémantiques a été rattaché à leur pouvoir discriminant dans la base, ce qui est lié au nombre de documents de la base où ils figurent.
En d'autres termes, la séparation suivant les caractères grammaticaux, entre mots sémantiques et mots fonctionnels, accompagnée de calculs statistiques de répartition sur la base, détermine des fonctions de poids qui expriment le degré de "signifiance" dans la base, des mots sémantiques.

Cette stratégie nécessitait l'élaboration d'un analyseur syntaxique qui après analyse morphologique, était capable d'identifier les catégories de tous les mots des textes de toute base et de toute question. Une fois faite l'identification du mot et de sa catégorie par une méthode de filtrage morpho-syntaxique, il est possible d'obtenir des relations de dépendances sémantiques.
Ces relations jouent un grand rôle dans la reconnaissance des polysémies, en identifiant la structure contextuelle des mots dans la phrase.

Il y a lieu de souligner que ces dépendances sémantiques sont obtenues automatiquement par un procédé de filtrage syntaxico-statistique. Sans faire appel à des dictionnaires de formes

composées préexistant, cette démarche, qui va de la reconnaissance grammaticale à celle des relations lexicales sémantiques, est inverse des analyses usuelles "Top-down".

III.4.1.3- Organisation algorithmique

Pour coordonner l'action des différents modules linguistiques et mathématiques, c'est-à-dire algorithmes et lois probabilistes, les systèmes SPIRIT et MICRO-MIND sont dotés d'une organisation qui comporte les modules suivants :

- ***Module 1:*** découpage du texte, analyse morphologique et reconnaissance des locutions, analyse syntaxique, sélection et lemmatisation des sémantiques.
- ***Module 2:*** création d'un fichier inverse des mots sémantiques lemmatisés, prise en compte de la dépendance lexicale sémantique, calcul automatique de la fonction de poids sémantique.
- ***Module 3:*** interrogation permettant de comparer une question donnée à tous les documents de la base, pondération du résultat de cette comparaison, obtention des réponses rangées dans un ordre décroissant de pertinence, renforcement de la qualité de la comparaison à l'aide des relations de dépendances.

III.4.1.3.1- Fichier inverse

Le fichier inverse permet l'accès direct aux documents d'une base à partir des constituants de ces documents (mots ou groupes de mots qui constituent les descripteurs). Il réalise une correspondance entre descripteur et document qui permet de savoir dans quels documents se trouve un descripteur donné de la base.

Selon la définition de l'auteur [ANDREEWSKY, 96], l'indexation automatique est la constitution du fichier inverse des mots sémantiques de la base, et la pondération de ces mots en fonction de leur pouvoir discriminant. Cette opération est entièrement automatisée.

La co-occurrence de deux termes (ou plus) dans un paragraphe ou dans une phrase ne peut être l'effet du hasard. La solution qui a été adoptée lorsque deux termes (ou plus) sont ainsi soudés et figurent sur une ligne du fichier inverse, est de former un tout, tant du point de vue conceptuel que celui des statistiques effectuées.

Le fichier inverse fournit aussi les pondérations des mots et groupes de mots afin d'optimiser la recherche des réponses à une question donnée, et de les ranger par ordre de pertinence. Nous traiterons ces points dans ce qui suit.

III.4.1.3.2- Segmentation du discours

La mise en place des traitements permet d'identifier les mots (unités lexicales) qui comparent les questions aux documents.
Ces unités peuvent être simples à déterminer, comme les chaînes de caractères entre deux blancs. Elles peuvent être complexes et faire appel à un traitement contextuel, comme "aujourd'hui", "C.E.E.", etc.

D'autres unités plus complexes, comme "mise en place", "faire école", etc., nécessitent l'emploi de dictionnaires et aussi parfois un traitement du contexte pour relier les termes d'une même expression.

Cependant, avec des traitements élémentaires, il est possible de faire les premiers pas d'une identification lexicale et poser ainsi les prémisses de traitements linguistiques plus élaborés.

III.4.1.3.3- Mots sémantiques et mots fonctionnels

La comparaison question-textes se fait par l'intermédiaire de mots et groupes de mots. Le fichier inverse donne en une seule lecture tous les documents contenant un mot donné. Il est réduit aux seuls mots dits sémantiques (substantifs, adjectifs, verbes, adverbes) que l'on sépare à l'aide de procédures automatiques des mots dits fonctionnels (conjonctions, prépositions, articles, pronoms, auxiliaires, etc.).

Cette opération limite l'expansion du fichier interrogeable par rapport à la base. Notons que la séparation entre les mots fonctionnels et les mots sémantiques se fait sur des critères syntaxiques. Cela ne signifie pas que les mots fonctionnels soient vides de sens. Dans la langue, la plupart ces mots ont une fonction « relatrice » et donc une action sémantique au traitement. Sinon, "*vivre+/avec/+ ses moyens*" et "*vivre+ /au dessus de/+ ses moyens*" auraient le même sens.

A la sélection des mots sémantiques, des relations entre mots sémantiques sont obtenues après l'analyse syntaxique (filtrage syntaxique).

Dans les systèmes SPIRIT et MICRO-MIND, l'indexation automatique prend en compte tous les mots sémantiques du texte, les équivalences sémantiques et les relations de dépendances lexicales. Des calculs statistiques permettent de réguler l'ensemble du processus d'indexation et d'interrogation.

La qualité de cette indexation est évaluée par la pertinence de l'ordre dans lequel les réponses sont fournies. La hiérarchie des réponses s'obtient à l'aide de fonctions de poids sémantiques (ces fonctions seront exposées ultèrieurement).

III.4.1.3.4- Analyse syntaxique

Pour séparer les fonctionnels des sémantiques, en français, l'identification des structures grammaticales des énoncés est nécessaire. Beaucoup de mots fonctionnels peuvent aussi être des mots sémantiques.

Ces mots ambigus entre fonctionnels et sémantiques, bien que peu nombreux dans la langue, se rencontrent très fréquemment dans le discours.

Dans MICRO-MIND, la séparation entre fonctionnels et sémantiques se fait à l'aide d'un analyseur syntaxique comprenant un millier de règles.
Dans SPIRIT, la reconnaissance des catégories grammaticales s'effectue grâce à l'emploi d'un analyseur syntaxique dont les règles sont obtenues par apprentissage à partir de textes désambiguïsés. Le modèle de l'analyseur est de type Markovien.

♣ Les dictionnaires

Le dictionnaire est constitué de 500 000 formes complètes du français et inclut les formes désaccentuées. Dans chacune de ces formes, on trouve la liste des valeurs grammaticales possibles hors contexte et éventuellement le genre et le nombre.

Exemple :
Le mot "ton" sera codé par : "article possessif" et "substantif masculin singulier".

Ce dictionnaire représente le savoir grammatico-lexical du système. Il sert d'abord à détecter les erreurs orthographiques de la base et les nouveaux mots par rapport au dictionnaire. Ensuite, les formes idiomatiques et les mots composés.

♣ L'analyseur

L'analyseur syntaxique va identifier automatiquement les catégories grammaticales correctes des énoncés et les analyses incorrectes seront écartées.

Exemple :
L'énoncé "ton car part !" sera analysé par :
- article possessif+substantif+verbe conjugué, sera retenue
et l'analyse
- substantif+conjonction+substantif, sera écartée.

Dans SPIRIT, les règles syntaxiques de l'analyseur sont construites à l'aide d'une méthode d'apprentissage. Ce qui implique au préalable l'étiquetage grammatical d'un corpus initial très diversifié. A l'aide d'une procédure d'auto-cohérence, l'étiquetage de ce corpus initial est confronté au même corpus où chaque mot est affecté de toutes ses catégories grammaticales possibles hors contexte. Ce qui permet de déterminer les règles de syntaxe « inférentes » pour analyser et désambiguïser un texte.

De tels analyseurs Markoviens à apprentissage réalisent une bonne séparation entre fonctionnels et sémantiques, ainsi qu'une bonne identification et lemmatisation éventuelle des verbes, adverbes, substantifs et adjectifs.

♣ Termes équivalents et termes parents

L'analyseur syntaxique permet aussi de traiter les homographes grammaticaux, ce qui facilite la recherche automatique :

1- ***des termes équivalents constitués :***

- les synonymies grammaticales sont le résultat d'une dérivation à partir d'une racine commune d'un groupe de mots de même catégorie grammaticale.
  Exemple :
  Livrerait=livrer=livre=livres.

- les synonymies conventionnelles sont des synonymes spécifiques à la base.
  Exemple :
  Cosmonaute=astronaute.

- Pour certains linguistes, la véritable synonymie n'existe pas dans la langue. Deux termes ne peuvent être considérés comme synonymes que dans le contexte d'une situation, d'un vécu, ou d'une base donnée.
  Exemple :
  Par rapport à un système de taxes, les deux termes *"péniche"* et *"remorqueur de rivière*" peuvent être des synonymes. Mais, par rapport aux constructeurs, ils ne le sont pas.

2- ***des termes dits parents :***

Ils se différencient par un suffixe, mais ils recouvrent le même champ sémantique (par rapport à la question).
Exemple :
Chanson=chansonnette.
Dans certains cas, on peut même accepter: *chansonnier=chanson.*

Le traitement de la parenté sémantique exige de faire appel soit à un thésaurus de termes parents, soit à une stratégie de séparation en "racines+terminaisons".

La dernière approche "racines+terminaisons" est moins figée et plus inférente, mais exige une validation pour éviter des équivalences incorrectes.
Exemple :
Voisé≠ voisin, bien que tous deux ont la même racine et sont équivalents par rapport aux terminaisons respectivement *« é »- « in » / « sé »- «sin ».*

Ainsi, on peut juger le rôle respectif et l'importance de la reconnaissance automatique des termes équivalents ou parents. Si l'on désire qu'une recherche documentaire donne ces termes équivalents, on doit :

- normaliser les formes conjuguées et les pluriels, ce qui après désambiguïsation grammaticale donne les équivalences en termes (équivalents constitués).
- une fois cette première équivalence faite, on peut, par une recherche dans le thésaurus, obtenir des termes parents.

Ce dernier type de synonymie peut demander un traitement complémentaire du contexte.
Par exemple : "travailleur manuel" ≠ "manuel du travailleur",
"livre (sterling)" ≠ "livre (d'étude)".

III.4.1.3.5- Rareté sémantique et calcul effectif de la rareté

♣ Définition

La rareté est une grandeur qui varie dans le sens inverse du nombre de documents dans lesquels une "forme" apparaît. Par cette forme est désignée, *l'unité de traitement désambiguïsé*, qui regroupe ses diverses formes synonymes.
La valeur numérique de la rareté d'une forme est l'inverse du nombre de documents dans lesquels cette forme apparaît. Plus ce nombre est faible, plus la rareté de la forme est importante.

Comment la rareté peut-elle être associée à une sémantique ?

Dans le cas limite où un mot figure qu'une seule fois dans une base, si cette base est interrogée avec ce seul mot, il n'y aura qu'un seul document réponse.
Autrement dit, la rareté est expliquée dans l'aptitude des mots et groupes de mots à discriminer des textes. L'expérience montre que jusqu'à un certain point, cette aptitude coïncide avec ce qui est significatif (sémantiquement important) pour l'utilisateur.

D'autre part, la rareté dans une base peut s'expliquer par des raisons autres que sémantiques. Comme le style de l'auteur, une maladresse terminologique, un mot mal orthographié, etc.

Lorsque l'utilisateur formule une question avec des mots ou groupes de mots peu distribués dans la base, la réponse est satisfaisante et il s'agit de documents les plus spécialisés de la question.

♣ **Calcul effectif de la rareté**

Le calcul de la rareté se fait de plusieurs manières: entropie, dispersion, nombre de documents dans lesquels le mot apparaît.

L'entropie et la dispersion font appel au calcul de la probabilité avec laquelle un mot sémantique M d'une base B caractérise chacun des documents D_i de B.
Cette probabilité $P(D_i /M)$ est d'autant plus grande que M est très occurrent dans D_i , et rare dans les documents D_j (différent de D_i) :

$$P(D_i / M) = \frac{P(M / D_i)}{\sum_{j \neq i} P(M / D_j)} \qquad \text{avec} \quad \sum_i P(D_i / M) = 1 \ ;$$

Si M ne se trouve que dans un seul document D_i , alors $P(D_i / M) = 1$.

Cette formule peut être obtenue d'une façon plus rigoureuse en assimilant le processus documentaire à un modèle de tirage exhaustif d'événements incompatibles.
Dans ce modèle, chaque document est une urne comprenant un ensemble de mots et la base est une urne comprenant un ensemble de documents.

D'où l'expression : $F(M) = \log N - H(M) + 1$

$$\text{avec} : 0 \leq H(M) = -\sum_i P(D_i / M) * \log P(D_i / M) \leq 1$$

H(M): définie l'entropie du mot M dans la base B des documents D_i .
F(M): est une fonction poids, définie de façon à ce que qu'elle soit d'autant plus grande que M est rare dans la base .

Lorsque la base est modifiée, une gestion simple permet d'obtenir la nouvelle valeur H(M) et donc F(M) à partir de son ancienne valeur, pour tous les mots concernés.

III.4.1.3.6- Interrogation et reformulation de la question

A l'aide des fonctions de poids sémantiques, le traitement d'une question donnée permet d'effectuer des intersections ensemblistes entre les mots sémantiques et le fichier inverse de la base.
Les mêmes procédures de traitements des documents sont utilisées pour la question et permettront la sélection des mots sémantiques et de leur adjacence.
Soit, une analyse morphologique, une analyse syntaxique et une sélection des mots sémantiques et des termes adjacents.

Dans SPIRIT et MICRO-MIND, on assiste à une comparaison, par l'intermédiaire du fichier inverse, entre textes et énoncé-question. De la même manière, il est possible de considérer n'importe quel texte-document comme celui de la question.
Il est possible de définir des listes de termes équivalents (termes synonymes, termes génériques-spécifiques, etc.) et de termes inférés. A partir de cette reformulation de termes, le système peut calculer une seule pondération à la classe d'équivalence et sans intervention de l'utilisateur.

III.4.1.4- Conclusion

Quatre modules imbriqués constituent les fondements des systèmes SPIRIT et MICRO-MIND :

- ***le module linguistique***
  La segmentation du texte, la désambiguïsation entre fonctionnels et sémantiques, la normalisation des sémantiques, et la détermination des dépendances de ces mots par des fonctionnels.
- ***le module probabiliste***
  La calcul de la fonction poids de chaque mot ou groupe de mots sémantiques, puis rangement hiérarchique des documents-réponse.
- ***le module du fichier inverse***
  A chaque mot sémantique de la base (outre son poids) est associé à la liste des documents où il figure. Ce module se réorganise à chaque modification de la base.
- ***le module d'interrogation***
  Ce module utilise les trois modules précédents, pour une intersection ensembliste pondérée entre mots sémantiques de la question et du fichier inverse.

III.4.2- PAPINS

Introduction

Le prototype PAPINS est le résultat d'une étude de faisabilité sur les techniques linguistiques applicables pour orienter le développement d'outils industriels.

Il est le résultat de modélisation et d'implémentation d'un outil expérimental d'extraction de connaissances à partir de textes qui sont basées sur des descriptions de la sémantique lexicale.

Ce prototype a pour objectif de fournir une représentation structurée du contenu des textes décrivant les projets de recherche de la DER (Direction des études et des recherches) de l'EDF (France) [PUGEAULT, 92].

III.4.2.1- Objectifs industriels d'EDF en matière d'indexation

L'objectif d'EDF en matière d'indexation est d'analyser globalement et massivement l'information présente dans les documents d'origine interne ou externe transitant à la DER.
Les techniques linguistiques sont ainsi utilisées pour structurer l'information textuelle, c'est-à-dire transformer le texte libre rédigé en termes d'indexation qui normalisent le contenu.
Les index sont exploités pour répondre aux différents besoins de l'entreprise :

- construction de fonds documentaires,
- classification et comparaison de documents,
- recherche et diffusion de l'information documentaire,
- veille technologique, etc.

III.4.2.2- Mise en oeuvre et évaluation d'un prototype

III.4.2.2.1- Organisation générale du système

L'objectif technique de PAPINS est d'extraire puis de représenter de manière structurée le contenu sémantique des textes (actions de recherche) en introduisant des données de la sémantique lexicale.

Les approches du traitement automatique du langage naturel (T.A.L.N.) à valider dans les analyses ont pour utilité :

- de montrer la faisabilité automatique de l'étude,
- d'évaluer le coût de mise à jour des ressources linguistiques,
- d'évaluer le rôle du lexique,
- d'évaluer les stratégies d'analyse,
- et de l'adéquation des résultats obtenues par rapport aux objectifs de l'EDF.

Le système se décompose en trois niveaux d'analyse successifs correspondant à trois niveaux de modules informatiques des ressources et des formalismes utilisés:

- ***Le niveau 1***
  consiste à structurer les textes de manière à en représenter l'organisation externe. Ce niveau est appelé le niveau d'analyse pragmatique.
- ***Le niveau 2***
  consiste en l'extraction de formes prédicatives. C'est le niveau d'analyse linguistique.
- ***Le niveau 3***
  décrit la phase de réécriture de prédicats selon un formalisme de représentation plus profonde. C'est le niveau conceptuel.

III.4.2.2.2- Spécifications techniques de PAPINS

Nous allons décrire les objectifs de chacun des niveaux du prototype et les points importants de sa mise en oeuvre :

♣ ***Le niveau 1 :***

Les ressources spécifiques qui sont requises à ce niveau sont les suivantes:

- un lexique : informations morpho-syntaxiques et sémantiques,
- une grammaire : qui sert d'interface entre le lexique et les règles d'extraction,
- une liste de déclencheurs : que sont des marqueurs linguistiques identifiés sur une analyse de corpus et permettant de cibler l'information à extraire.

La modélisation de l'extraction de fragments de phrases associés à une articulation se réalise par le biais d'une base de règles d'extraction mettant en oeuvre les déclencheurs (70 règles d'extraction) .
Le niveau 1 reçoit en entrée un ensemble de textes pré-traités mis en format Prolog ou les ARD (textes techniques complexes de 300-400 mots par texte).
En sortie, il produit pour chaque ARD des fragments de phrases pertinents qui sont organisés en articulation (thème, motivations, problèmes et réalisations).

♣ ***Le niveau 2 :***

Les ressources linguistiques requises au niveau 2 de PAPINS sont :

- le lexique,
- les classes sémantiques de verbes qui caractérisent les prédicats,
- les types sémantiques de prépositions,
- le thésaurus EDF qui fournit une hiérarchie de descripteurs, de thèmes et de champs sémantiques,
- les restrictions de sélection sur la nature des syntagmes nominaux (les arguments),

- une base de rôles pour constituer les grilles thématiques des prédicats,
- un analyseur morpho-syntaxique.

La modélisation du niveau_2 se réalise par la formalisation d'une base de critères (assignation de rôles thématiques). Ces critères permettent d'assigner un rôle thématique à un argument d'un prédicat en considérant sa classe et le type sémantique. le type est un syntagme prépositionnel.

Le niveau_2 reçoit en entrée les fragments de phrases pertinents organisés en articulations pour chaque ARD traitée.

Pour chaque texte d'ARD, Les sorties produites sont des fragments de phrases pertinents organisés en articulations et représentés sous la forme de prédicats-arguments et prédicats-modifieurs.

Une structure prédicative est de la forme (prédicat(rôle thématique, argument)*)*.

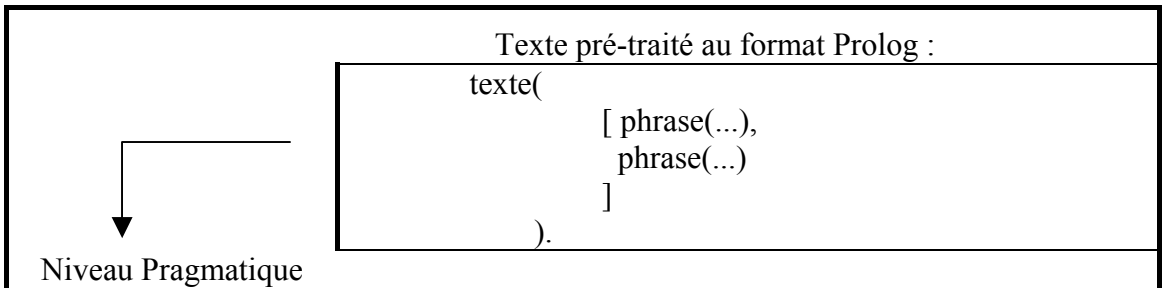
♣ ***Le niveau 3 :***

L'objectif du niveau 3 est d'améliorer la qualité de l'exploitation des résultats du niveau_2, c'est-à-dire représenter les formes prédicatives sous un format plus génériques.
Les entrées correspondent à certaines classes de verbes du corpus d'EDF qui décrivent des prédicats dénotant des actions complexes et méritant d'être décomposés en entités conceptuelles.
En sortie, on retrouvera les représentations de chaque classe selon un format type (LCS).
La modélisation du niveau conceptuel se réalise par la définition de primitives, de champs sémantiques et de catégories conceptuelles qui permettent de représenter les classes de prédicats du corpus EDF.

III.4.2.3- Conclusion

D'un point de vue linguistique, les méthodes mises en oeuvre dans le prototype PAPINS sont bien adaptées aux problèmes exposés par l'EDF.

Les trois niveaux de représentation de connaissances sont homogènes et sont exprimés selon un formalisme incrémental (*Fig.III.4.2.3.*). Des méthodes sont définies pour extraire des termes pertinents notamment en utilisant un thésaurus.



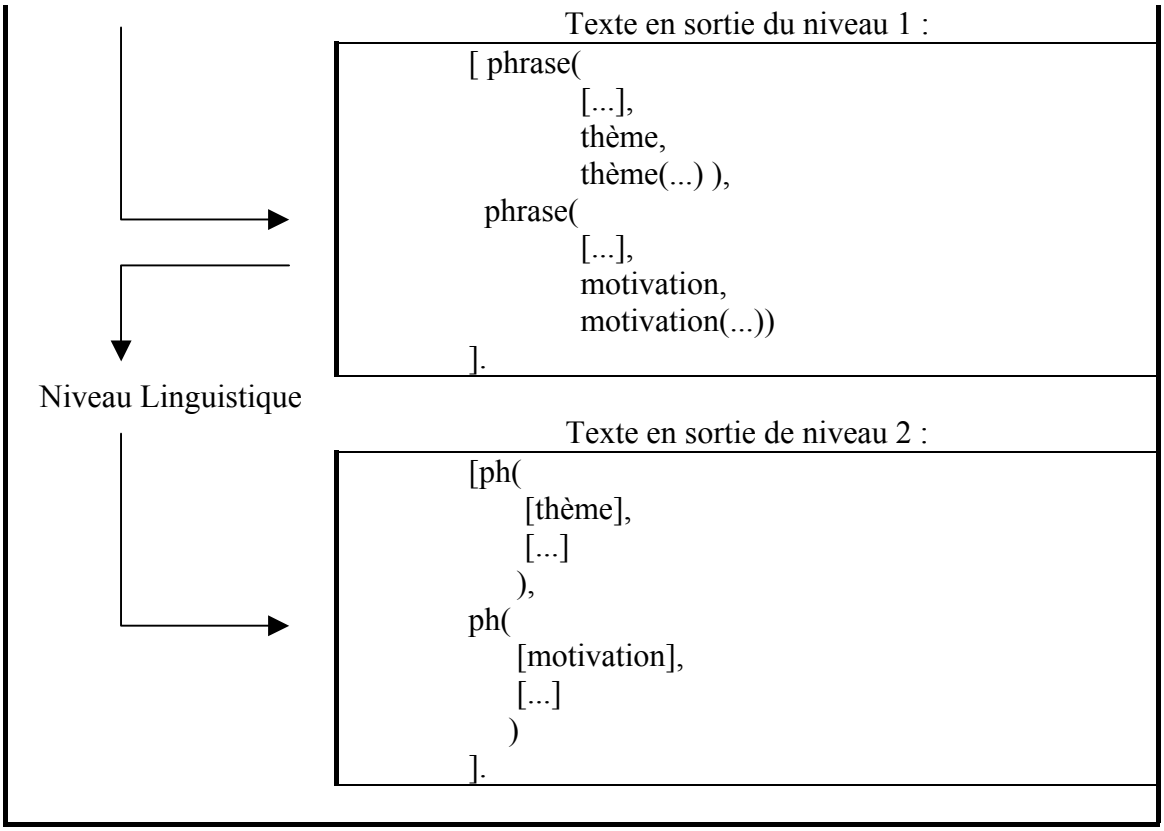


Fig.III.4.2.3. Système de pré-traitement, le niveau 2 et le niveau 3.

III.4.3- TERMINO

III.4.3.1- Introduction

Termino est un logiciel de dépouillement terminologique assisté par ordinateur. Il a été conçu et réalisé par le groupe Recherche de Développement en Linguistique Computationnelle (R.D.L.C.) du centre en analyse de texte par ordinateur (A.T.O.) de l'Université du Québec à Montréal (U.Q.A.M.), S. David et P. Plante (1990).

La méthode d'indexation automatique proposée est basée sur une analyse syntaxique de texte combinée à une analyse statistique de pondération des termes [PLANTE, 88].

Cette approche explore deux voies et repose sur la combinaison de méthodes statistiques traditionnelles et d'analyse syntaxique pour des textes français.

Les deux voies considérées pour améliorer ces méthodes :

- ***La génération de termes composés***
  Les termes composés permettent généralement de limiter l'ambiguïté des termes et d'augmenter la précision.
- ***L'analyse de la langue naturelle***
  une autre voie intéressante pour l'amélioration des performances de l'indexation automatique est de tenter de tirer profit des méthodes d'analyse de la langue naturelle. Dans le but de raffiner le processus d'indexation, la levée des ambiguïtés de lemmatisation permet de suggérer des termes composés.

III.4.3.2- Méthode d'indexation automatique

Dans ce projet, l'analyse faite par le logiciel Termino est le point de départ du processus d'indexation. L'analyse syntaxique produite par Termino permet, entre autre, de lever diverses ambiguïtés de lemmatisation et de générer des termes composés à partir d'une analyse syntaxique poussée.

A partir de la sortie de Termino, un logiciel développé produit une indexation automatique en appliquant des méthodes statistiques. Le logiciel est paramétrisé afin de pouvoir étudier diverses combinaisons possibles au niveau du choix des termes et des méthodes de pondération.

III.4.3.2.1- Aspects linguistiques et informatiques de Termino

La théorie syntaxique mise en oeuvre dans Termino est une théorie positionnelle [PERRON, 89a]. L'hypothèse fondamentale est la suivante : l'analyse syntaxique d'une phrase n'est pas entièrement déductible des unités lexicales qui la constituent. On distingue dans ce modèle :

- un ensemble de position, c'est-à-dire un ensemble de « points » hiérarchiquement organisés entre eux pour former une *configuration*.
- une relation entre les positions et les constituants lexicaux, appelée *relation d'occupation.*

L'objet de la syntaxe est alors de rendre compte des positions (identification des positions et de leurs propriétés spécifiques) et de l'occupation de ces positions. Les éléments lexicaux relèvent du composant lexical.
Termino est constitué de plusieurs modules :

- EDITO : traitement des marques d'édition,
- LCMF : lemmatisation et caractérisation morphologique du français,
- ALSF : analyseur lexico-syntaxique du français
- MRSF : Module de reconnaissance des synapsies du français,
- MRAF : module de rédaction assistée des fiches,
- FX : langage de programmation en faisceaux.

Lorsque l'opération « Description Linguistique » est lancée, les textes sont analysés et Termino produit une série de fichiers dont les principaux sont :

- le fichier de adjectifs,
- le fichier des noms,
- le fichier des verbes,
- le fichier des synapses.

Chacun des fichiers précédents contient un terme par ligne. Chaque terme est suivi par les numéros de phrases où il apparaît.

III.4.3.2.2- Caractéristiques des modules

♣ **EDITO**

Le module EDITO effectue le découpage du texte en phrases et en mots, et reconnaît les noms propres.

♣ **LCMF**

Pour chaque mot, le module LCMF fournit une catégorie grammaticale (nom commun, adjectif, verbe infinitif, verbe fléchi, participe présent, participe passé) et une caractérisation morphologique (traits de genre, nombre, personne, mode et temps selon la catégorie).
Si le mot présente une ambiguïté catégorielle, alors LCMF fournira toutes les hypothèses et la désambiguïsation s'opérera par le module ALSF.
Enfin, LCMF fournit pour chaque mot son lemme, pour :

- les noms commun : le singulier,
- les adjectifs : le masculin singulier,
- les verbes fléchis : l'infinitif,
- les participes présents ou passés : l'infinitif.

LCMF est essentiellement fondé sur une analyse morphologique des formes lexicales et il a la possibilité de catégoriser et d'assigner un lemme à des néologismes.

♣ **ALSF**

Ce module fournit, à l'aide d'une analyse descendante, une représentation hiérarchisée des différents groupes syntagmatiques de la phrase (groupe nominal GN/ verbal GV / prépositionnel GP/ adjectival GA, etc.) et la représentation des relations entre ces groupes.
ALSF est un « parseur » fondé sur une théorie syntaxique positionnelle, il s'appuie à la fois sur un savoir syntaxique (la géométrie des positions) et sur un savoir lexical (le sous-ensemble des propriétés lexicales et leur mise en jeu dans tel ou tel position).
De ce fait, il est un parseur syntaxique autonome et sa représentation syntaxique est sous-déterminé (il ne construit pas toutes les possibilités de structuration ou de rattachement, etc.).

Le langage avec lequel est programmé ALSF est le langage FX qui est bien adapté à la manipulation des positions dans un arbre.

♣ **MRSF**

Le module MRSF est chargé de construire des synapsies. Une synapsie est une unité nominale polylexicale (U.N.P.), c'est-à-dire formée de plusieurs termes, construite syntaxiquement. Elle est composée de groupes GP, GN et GA. La tête (ou base) de la synapsie T est un nom commun, et représente son noyau (*Fig. III.4.3.2.2.*).

Exemple:

Forme	Exemple
T+GA	plante verte
GA+T	haute tension
T+GP	traitement de textes
T+GA+GA	horaire décalé fixe
T+GA+GP	logiciel intégré de gestion
T+GP+GA	système de gestion universel
T+GP+GP	valeur de l'actif par action

GA+T+GP	grand livre de banque
GA+T+GA	grand livre auxiliaire
T+GA+GP+GP	nombre moyen de défauts par unité
où : T: tête GN/GP/GA: groupe nominal/prépositionnel/adjectival.	

Fig. III.4.3.2.2. Exemples de synapsies.

La détermination des synapsies consiste en une exploration de la représentation syntaxique fournie par ALSF. A chaque synapsie est associée une représentation qui met en jeu à la fois sa structuration syntaxique et une trace de son ancrage syntaxique dans la phrase.

Parmi les quatre catégories de termes produites par Termino, toutes les catégories ou une combinaison de catégories sont choisies pour l'indexation [PERRON, 89b]. Les termes composés correspondent aux synapsies produites par Termino. Chaque combinaison est identifiée par un code (wxyz) de quatre lettres indiquant quelles catégories de termes sont utilisées (*Fig. III.4.3.2.2b.*).

Le code = (wxyz) où : w,x,y,z $\in \{S, A, V, N\}$, sont des catégories de termes, S : synapsies, A: adjectifs, V: verbes, N: noms.
Exemple : Le code SxVN : représente la combinaison des synapsies (S), verbes (V) et adjectifs(A). Le x indique que cette catégorie a été omise. En tout, il y a 15 combinaison possible de ce code.

Fig. III.4.3.2.2b. Constructions de synapsies.

III.4.3.2.3- Méthode de pondération statistique

Il existe plusieurs méthodes pour calculer l'importance (ou poids) d'un terme simple dans un document. L'approche qui a été utilisée est celle décrite par les chercheurs Salton G. et Buckley C. en 1988 [SALTON, 89].

Dans cette approche, le poids d'un terme simple est déterminé à partir du produit de trois composantes :

- sa fréquence dans le document (C1),
- sa fréquence dans la collection de documents (C2),
- et un facteur de normalisation (C3).

C1 : fréquence du terme		
b	1 ou 0	poids binaire égal à 1 si le terme est présent
t	ft	fréquence du terme
n	½+½ (ft/max ft)	fréquence normalisée augmentée
C2 : fréquence dans la collection		
x	1	aucun changement
f	logN/fd	fréquence dans la collection inverse, N= nombre de documents, fd= nombre de documents auxquels le terme est associé.
p	log(N-fd/fd)	fréquence dans la collection inverse probabiliste
C3 : facteur de normalisation		
x	1	aucune normalisation
c	$\frac{1}{\sqrt{\sum_{i(termes)} p_i^2}}$	normalisation

Tab. III.4.3.2.2. Composantes du poids d'un terme simple [SALTON, 89], [FARAJ, 96].

Ainsi, le tableau (*Tab.III.4.3.2.2*) représente les méthodes de calcul de ces trois composantes (18 combinaisons possibles). Chaque combinaison sera représentée à l'aide de trois lettres où chaque lettre représente le paramètre utilisé pour chacune des composantes.

La méthode de calcul de poids (tfx) signifie que pour le composant C1 le paramètre (t) a été utilisé, que pour le composant C2 de paramètre (f), et pour le composant C3 de paramètre (x).

En ce qui concerne les termes composés, la détermination du poids est une question qui n'a pas encore de réponse précise. La méthode retenue est celle de Croft, Turtle et Lewis (1991) [CROFT, 91], que le terme composé est comme le terme simple et d'appliquer les mêmes méthodes de pondération [TURTLE, 91].

III.4.3.3- Modèle de repérage requête-documents

Le modèle vectoriel est utilisé pour Termino. Dans le modèle vectoriel chaque document i indexé est représenté à l'aide d'un vecteur de la forme :

$$D_i = (p_{i1}, p_{i2}, ..., p_{it})$$

où

p_{ik} : représente le poids du terme k dans le document i,

t : représente le nombre total de termes dans la collection.

Pour faire une recherche, l'utilisateur soumet sa requête j en langue naturelle au système. Ce dernier analyse la requête et calcule le poids des termes r_{jk} de la requête.

Ainsi, Termino représente la requête à l'aide d'un vecteur R_j : $R_j = (r_{j1}, r_{j2}, ..., r_{jt})$

Les différentes méthodes de pondération décrites à la section précédente peuvent être appliquées non seulement aux documents mais aussi aux requêtes. Le code (D/bxx) représente le fait que la méthode de pondération (bxx) a été utilisée pour les documents (D) et le code (R/bxx) pour la requête (R).

Le système calcule un coefficient de similarité entre la requête et les documents. La formule choisie pour calculer ce coefficient est la suivante :

$$sim(D_i, R_j) = \frac{\sum_{k=1}^{t} p_{ik}^2 * \sum_{k=1}^{t} f_{jk}^2}{\sqrt{\sum_{k=1}^{t} p_{ik}^2 * \sum_{k=1}^{t} r_{jk}^2}}$$

Après avoir calculé ce coefficient pour tous les documents, le système trie les documents par ordre décroissant par rapport à ce coefficient et présente à l'utilisateur cette liste triée. De ce calcul, les premiers documents présentés sont les plus similaires à la requête.

III.4.3.4- Conclusion

Dans le cas de la présente collection de documents testée par Termino, l'utilisation de termes composés générés par analyse syntaxique produit une amélioration systématique de la performance par rapport aux termes simples. Ces résultats contrastent avec d'autres études du même genre sur la langue anglaise [FAGAN, 87].

Plusieurs explications sont possibles, surtout un facteur à considérer est la plus grande richesse syntaxique de la langue française par rapport à la langue anglaise [AMAR, 92].

D'autres variations dans la méthode de pondération des termes composés sont aussi à considérer [OTMAN, 92].

III.4.4- LEXIC

Introduction

Lexic (Système de Recherche d'Information en Langage Naturel, de GEDIMAGE France, version 2.12 de 1995) est l'aboutissement de plusieurs années de recherche menées conjointement dans les domaines de la linguistique informatique et des applications liées au traitement des bases de données non-structurées.

Lexic est un moteur d'application dans le domaine de la gestion électronique de l'information et de la documentation. Il permet d'accéder à des documents textuels en utilisant la *valeur sémantique* de leurs contenus [GEDIMAGE, 95].

III.4.4.1- Indexation lexicale automatique

A partir d'une méthode originale basée sur la simulation de réseaux neuronaux, Lexic offre la possibilité de rechercher un élément du texte par simple interrogation en langage naturel.

Il est directement opérationnel sur les documents textuels en format ASCII grâce à son indexation lexicale automatique.

L'utilisateur a la possibilité d'associer un texte à un document graphique et inversement dans l'hypothèse où ce dernier est l'objet de la recherche. Ainsi, l'exploitation conjointe des textes et de l'image illustre davantage les documents.

III.4.4.1.1- Lexic et informatique linguistique

Lexic s'oriente vers l'analyse linguistique afin de comprendre le sens des textes. L'interprétation d'une phrase oblige à poser des hypothèses et d'en dégager les outils de travail nécessaires. Ce travail nécessite des techniques et des méthodes pour aller vers une optique purement linguistique. Il propose une assistance automatisée originale à la gestion de l'indexation et du thésaurus:

- ***au niveau lexical :*** il effectue la reconnaissance automatique des synonymes et ses expressions équivalentes,
- ***au niveau du sens du texte :*** il effectue le repérage des analogies, des voisinages de sens et des structures sémantiques.

Il constitue de la sorte une assistance à l'identification des termes associés et génériques, et à la définition des champs sémantiques, enfin, la détermination d'un concept relevant de l'intégralité du corpus.

La philosophie suivie pour la solution linguistique est que :

> *Lexic doit être un outil riche, flexible pour l'analyse logique et l'investigation sémantique du texte.*

Une intégration de Lexic aux différents environnements informatiques, grâce à :

- sa portabilité sur les systèmes d'exploitation : Windows 3.x / 9.x / NT , OS/2.
- son environnement ergonomique et convivial : Classement virtuel graphique (WinSide), des grilles de saisie allant de la plus simple aux plus complexes (graphes arborescents)
- son interaction dans les architectures en réseau : Novell, Lan Manager, Windows NT, etc.
- sa technique de compression de textes et des images : qui réduit la taille des fichiers et accélère la recherche,
- ses outils de développement : les API
- son intégration aux systèmes de gestion de bases de données relationnelles.

III.4.4.1.2- Lexic et recherche en texte intégral

D'un point de vue conceptuel, Lexic est un système de recherche en texte intégral simulant un réseau neuronal. Ce système est doté d'une interface graphique pour intégrer les fonctions qui permettent une exploitation continu du système.

le problème de la recherche en texte intégral soulève avant tout un problème de représentation agrégée des informations textuelles :

> *« Les systèmes de gestion de bases de données éprouvent de sérieuses difficultés à traiter les données qui sont dites non structurées : les données qui ont une signification non pas par leurs structurations mais par leurs relations logiques et sémantiques. Les bases de données limitent ainsi leur analyse à un traitement de style syntaxique ou de niveau linguistique supérieur. »,* ([GEDIMAGE, 96], p.2).

L'originalité du système Lexic est de reposer sur une représentation des textes d'un point de vue sémantique. La codification des textes, phase appelée compilation, repose sur un système de descriptions formelles qui garantissent la cohérence et l'unicité de la représentation.

III.4.4.2- La codification des textes

La codification des informations textuelles repose avant tout sur une analyse sémantique des mots et des phrases. Cette codification dépend de la structure grammaticale des phrases étudiées : *position des mots pris dans le contexte d'une phrase*.

Une telle démarche a été développée selon une approche linguistique et non informatique, car l'objectif est d'aboutir à une représentation symbolique des textes sans indexation. Pour cela, le symbolisme s'appuie sur une représentation codée par des règles formelles. Cette représentation théorique est valable pour des textes multilingues. Pour le moment seuls les langues française et anglaise sont supportées.

Lexic met en oeuvre dans cette représentation un système qui pourra être vu comme un "mémoire" de la connaissance contenue dans les textes :

> un dictionnaire de mots (formes) et des règles grammaticales permet d'améliorer l'efficacité des recherches.

Le dictionnaire de formes comporte 300,000 mots qui sont porteurs des informations suivantes :

- racines étymologiques pour 85% des mots : racines grecque et latine,
- synonymes : sens concret ou abstrait,
- pondération phraséologique : en fonction du domaine d'application, certains mots peuvent avoir soit une signification autre que celle communément admise, soit un poids dont il faudra tenir compte.

Dans certains cas, un dictionnaire de jargon pourra venir compléter le dictionnaire de formes. Son rôle est de rendre compte du contexte d'utilisation des textes, des formes particulières d'expression comme les abréviations, ainsi que des équivalences logiques (synonymes, raccourcis d'écriture, etc.).

III.4.4.3- Interrogation en langage naturel

L'utilisation du langage naturel comme moyen de recherche est une conséquence logique du travail d'analyse effectué sur les textes. Le formalisme de codification a été construit de telle manière qu'une interrogation de ce style soit naturelle et exclut la mise en place d'un traitement particulier (dictionnaire de mots, thésaurus).

Une question en langage naturelle est codifiée de la même manière que le texte. La question subit le même processus d'analyse et de codification que celle pour indexer le texte d'un document.

L'identification des réponses repose sur un processus de mise en correspondance entre la représentation symbolique de la question et celle d'un texte. Suivant le mode de recherche, précis ou étendu, l'identification pertinente des réponses est effectuée en terme d'égalité de sens : strict, approché ou élargi.

La mise en correspondance de la question posée avec les phrases du texte est effectuée par la simulation d'un ***réseau neuronal***. Par ce biais, les réponses sont nuancées car le processus d'identification ne repose pas sur une logique formelle et déterministe.

III.4.4.4- Conclusion

Lexic exécute en premier un traitement à partir de l'analyse morphologique des mots (genre, nombre, conjugaison,…), de la reconnaissance des synonymes, des étymologies et des voisinages de sens.

Il permet l'interrogation directe d'un fonds documentaire et la formulation de la requête en langage libre, sans aucune contrainte syntaxique (booléenne et arithmétique) ni référence à un type d'indexation (liste de descripteurs ou thésaurus).

III.4.5- THAÏS

Introduction

Le critère de la modernité dans les méthodes de conception des systèmes d'information documentaire réside essentiellement dans l'adoption des traitements aux problèmes posés.

Par ailleurs, la voie empruntée par l'équipe SYDO-Lyon, aux problèmes posés par l'indexation automatique de documents textuels, s'est avérée extrêmement difficile et éminemment linguistique : l'analyse automatique de la langue naturelle.

Ceci dit, une bonne partie de ce travail concerne des aspects linguistiques dont le modèle est attribué aux auteurs A. Berrendonner, M. Le Guern, R. Bouché, J.-P. Metzger et certains autres chercheurs.

Une première approche linguistique et informatique de ce modèle a été proposée dans la thèse de Marcilio De Brito [DE BRITO, 91]. Ce dernier ayant travaillé sur la conception d'un système prototype THAÏS pour l'indexation automatique de documents textuels. le corpus choisi ayant été composé de dépêches provenant de l'Agence France Presse.

L'analyse morpho-syntaxique de Thaïs est issue de la grammaire du Syntagme Nominal (SN). Cette étude fait l'objet de nombreux travaux de recherche du groupe SYDO-Lyon qui a beaucoup évolué dans sa conception.
THAÏS issu de ce modèle linguistique et des réflexions pratiques a été incorporé dans un programme informatique pour pouvoir démontrer la cohérence de ces règles et mettre en oeuvre les aspects évolutifs de ce projet.

III.4.5.1- Le contexte du Projet THAÏS

La réalisation d'un travail comme celui d'un analyseur morpho-syntaxique du français avec un compilateur (STARLET) en cours de test est confronté à trois difficultés :

- la transcription du modèle linguistique,
- le comportement du langage STARLET, et
- la conception de l'outil informatique lui-même.

Le contexte de coopération entre équipe de linguistes et celle des informaticiens était donc marqué par des efforts intellectuels pour remédier aux erreurs de l'implémentation qui pourraient être dues à :

- une mauvaise représentation d'un phénomène linguistique par la grammaire du SN,
- une erreur dans le compilateur STARLET,
- une erreur de programmation.

L'évolution conjointe du modèle linguistique (l'analyseur) et le langage de programmation qui le reçoit (le compilateur STARLET).

III.4.5.2- L'environnement de l'analyse morpho-syntaxique

L'analyseur est destiné à opérer avec une grammaire particulière gouvernée par la nature du corpus d'entrée, une série de dépêches de l'AFP. La grammaire du français conçue est spécialisée dans le traitement d'un corpus physiquement non limité mais linguistiquement restreint, car il était vain de penser à un analyseur universel.

Par ailleurs, la grammaire employée est celle d'un système orthographique français, ce qui entraîne tantôt des avantages tantôt des inconvénients : les régularités de l'écrit par rapport à l'oral sont parfois plus amples et les irrégularités originales.

Des spécifications qui ne font pas partie du corpus traité, comme :

- les formules scientifiques (chimiques, mathématiques, etc.),
- certaines homographies gênantes dues aux majuscules,
- etc.

III.4.5.2.1- Le corpus

Le texte qui a alimenté l'analyse comprend un corpus d'une série de dépêches de l'AFP dont la taille moyenne d'une dépêche est d'environ 300 mots. La diversité des auteurs et la variété des sujets traités fait de ce corpus un élément de caractéristiques hétérogènes.

La structure d'une dépêche se compose essentiellement de trois groupes d'informations :

- ***l'en-tête***
  qui contient toutes les informations susceptibles de définir et de reconnaître une dépêche.
- ***le corps***
  une dépêche est constitué généralement de trois parties: une ligne indiquant le titre de l'événement, un paragraphe représentatif du résumé de l'événement et le contenu textuel détaillé de l'événement.
- ***le descripteur***
  il contient une marque de fin de texte, un horodate et une chaîne de caractères indiquant le mois de l'année de la dépêche.

III.4.5.2.2- La phase morphologique

L'analyse morphologique doit fournir les données nécessaires aux composants ultérieurs :

- le module d'analyse syntaxique,
- le module d'indexation automatique,
- etc.

Pour répondre à ces exigences, la grammaire du SN est adaptée aux besoins de la phase d'analyse syntaxique :

> *« ... que le résultat de l'analyse morphologique contienne un maximum de structures régulières et pour ce faire nous ne devons pas nous contenter d'étiqueter les formes en surface du texte par des traits méta-linguistiques mais exécuter ces formes des*

opérations de régularisations, consistant à ramener les exceptions au cas général correspondant. », [DE BRITO, 91], p.99.

En se tenant au principe de réduction du complexe au simple, une opération de régularisation consistera à ramener une forme de surface à une forme profonde dotée d'un rendement maximal : la réduction du nombre de règles d'analyse (*Fig. III.4.5.2.2.*).

Exemple :

Cas général :	/au/→/à/$_{\text{préposition}}$ +/le/$_{\text{prédéterminant}}$
Exception :	/le/→/le/ $_{\text{prédéterminant}}$ ou /le/→/le/ $_{\text{pronom}}$

Fig. III.4.5.2.2. Exemple de schéma de représentation de forme.

III.4.5.2.3- Les ambiguïtés

Une importance particulière a été attachée aux phénomènes d'ambiguïtés, car s'est un point critique pour l'analyse automatique.

A un mot peuvent être associés un ou plusieurs ensembles de traits dans la phase d'analyse morphologique. Comme il est possible qu'aucun trait n'y soit affecté (c.f. la forme /ne/ étant une marque de négation, son complément /pas/ ne reçoit pas nécessairement de trait).

Ainsi, un mot est marqué comme suit :

- par l'affixe « INDEF », pour un mot non analysé ;
- par un seul ensemble de traits, pour un mot non-ambigu ;
- par un affixe composé représentant plusieurs ensembles de traits, pour un mot ambigu .

Exemple :

Mot	Statut	Traits
/parisyllabique/	mot inconnu	INDEF
/chaise/	mot connu	[F-NOM,FEM,SIN]
/certain/	mot ambigu	[F-ADJ] ou [D-INDEF]

Dans la majorités des cas, les ambiguïtés seront traitées par l'analyse syntaxique. Il n'existe pas un modèle de calcul pour les solutions "ambiguës", mais c'est le langage STARLET qui cherchera à valider les affixes imposés par un processus qui lui est propre : la confrontation.

III.4.5.2.4- Le processus de confrontation STARLET

Le passage des paramètres s'accompagne de tests de type et de conversions par confrontation (transmission du contexte = confrontation). Des arguments et des paramètres formels permettent de déterminer s'ils sont compatibles [DE BRITO, 92].

Cette confrontation se déroule de la manière suivante :

- Les paramètres formels en entrée définissent des contraintes de type sur le contexte d'entrée. A l'essai d'une règle, les valeurs des arguments doivent être compatibles avec les types des paramètres (un test à priori).

- Les valeurs calculées pour les affixes en sortie doivent être compatibles avec les types des arguments correspondants (un test à posteriori).
- La valeur d'une séquence d'affixes "s" est compatible avec un affixe "a" s'il existe des règles d'affixes telles que "s" dérive de "a". La valeur affectée à "a" sera composé" suivant ces règles. Etc.

Dans le cas d'une confrontation d'arbre, chaque valeur d'affixe est représentée par l'arbre syntaxique qui reflète la manière dont cette valeur (phrase d'affixe) a été construite. La confrontation peut s'effectuer sur un nombre quelconque de niveaux de l'arbre et la structuration est calculée d'après les règles d'affixes.

III.4.5.3- Le processus général : l'analyseur morpho-syntaxique

L'analyse morpho-syntaxique se déroule sur deux niveaux : dans l'un, une consultation directe du lexique et dans l'autre, un prétraitement morpho-syntaxique.

III.4.5.3.1- Le lexique

Le premier lexique conçu par A. EYANGO-MOUEN (Lexique interactif pour l'analyse automatique du français), est fondé sur le principe des "profils" lexical, syntaxique et flexionnel. Cette initiative intéressante pose certains problèmes : le modèle n'est pas ouvert à l'insertion automatique des mots dans le lexique.

Donc, envisager l'acquisition interactive automatique du lexique par insertion automatique par le biais d'une interface couplée à d'autres outils logiciels. Pour ce qui est des mots composés, la méthode la plus souple est l'intégration des expressions dans le lexique.

♣ **Lexiques : la structure STARLET**

Le lexique qui a été créé est dans le seul but de faire fonctionner l'analyseur. Il présente de nombreuses imperfections dont la solution réside dans le modèle d'acquisition interactive automatique du lexique.

Les caractéristiques des lexiques mis en place représentent les besoins immédiats en information. Les en-têtes de notation rappellent les spécificités de chaque lexique : la catégorie concernée, le corps de chaque en-tête comprend les profils lexical, flexionnel et syntaxique (*Fig. III.4.5.3.1*).

!<Nom-Catégorie>**(:**"<mot>"***,**"/<flexion>"***,** <traits catégoriels>**:):**[valeur de vérité**.**]

Fig. III.4.5.3.1. Forme d'entrée lexicale selon la structure STARLET.

Les structures de lexiques utilisées sont :
- lexique de pronoms pré-verbaux (la catégorie Y),
- lexique de formes en /que/ ,
- lexique de prépositions,
- lexique de noms-pronoms,
- lexique des adverbes,
- lexique des nominaux,
- lexique des prédéterminants,

- lexique des démonstratifs,
- lexique de locatifs,
- lexique des relatifs composés,
- lexique de subordonnants,
- lexique de coordonnants,
- lexique de mots-négatifs,
- lexique des mots particuliers.

Ces lexiques constituent une expression immédiate des besoins de l'analyseur pour mener à bien l'analyse en surface. Il est possible d'envisager par cette méthode "lourde" des nouvelles structures et l'insertion des nouvelles variables pour exprimer les liens logiques entre les formes du discours.

III.4.5.3.2- Le prétraitement morpho-syntaxique

Les aspects de l'analyse morphologique spécifiques au langage STARLET aboutissent aux résultats déjà annoncés par J-P Metzger [METZGER, 88] :

> Un prétraitement local de nature morpho-syntaxique précède brièvement l'analyse morphologique pour détecter, dans les séquences de formes, une propriété syntaxique quelconque. Ce traitement permettra la levée de certaines ambiguïtés par affectation "contextuelle" des catégories aux formes de surfaces.
> C'est aussi dans cette phase que l'on procède à la substitution des amalgames. Chaque amalgame sera ainsi remplacée par la suite de formes qui lui correspond. *«les effets de cette **opération de régularisation** du texte a pour conséquences une réduction de l'inventaire de formes du lexique ainsi que des catégories nécessaires à l'analyse»*.[DE BRITO, 91], p.114.

La reconnaissance d'une forme s'effectue, dans le cas le plus général, par l'application de règles d'analyse qui sont proposées suivant un ordre préétabli, en fonction des traitements linguistiques prioritaires. La complexité du traitement informatique joue aussi un rôle déterminant dans l'établissement des règles.

Le traitement morpho-syntaxique auquel M. De Brito a fait illusion est le modèle déjà annoncé par A. BERRENDONNER (1984) et J-P METZGER (1988). Ce traitement constitue une série de modèles répertoriant tous les comportements flexionnels possibles. A l'aide de ces modèles flexionnels, il est possible de réduire une forme fléchie à une forme canonique répertoriée dans le lexique.

Le mécanisme mis en oeuvre permet de réduire un mot fléchit (genre et nombre) comme /blanches/ à sa base /blanc/ (masculin singulier). A la forme résultante /blanc/ est affectée les interprétations linguistiques renvoyées par le lexique ou déduites par les règles d'analyse. L'ensemble de ces informations se compose (exemple : /blanches/) :

- d'une ***catégorie grammaticale*** (exemple: F pour les noms et adjectifs),
- d'une ***sous-catégorie*** (exemple : NOM pour les noms propres et communs),
- des ***valeurs flexionnelles*** en genre et en nombre(exemple : FEM, PLU).

III.4.5.3.3- Le traitement morphologique

Le traitement STARLET correspond à la définition d'une méta-grammaire (décrite par la grammaire d'affixes). Les définitions d'affixes seraient en programmation traditionnelle une déclaration de type.

La définition de **isfl** (**i**nformations **s**yntaxiques, **f**lexionnelles et **l**exicales) représente la suite de traits méta-grammaticaux affectés à chaque forme :

- **l** représente le mot analysé,
- **is** est une série d'informations ayant un comportement régulier. Chaque i peut recevoir une valeur de catégorie (categ) donc une valeur syntaxique, de sous-catégorie (scategs), une valeur flexionnelle (flexs), ou une valeur lexicale (lexs) :

categ : F; D; W; Y; Q; P; C; INDEF.

scategs : scateg; scateg scategs; vide.
scateg : NOM; NAN; ADJ;
DEF; NUM; IND;
COM; PRP; PRO;
INu; INd; INN; INda; INdb; INdn; INdo;
AAJ; QUA; PRO; TAM;
CI; LA; DE.

flexs : genre nombre; genre nombre personne.
personne : PEu; PEd; PEt; PEq; PEc.
nombre : SIN; PLU; NBN.
genre : MAS; FEM; GRN.

lexs : lex; lex lexs; vide.
lex : rection; négativité; contdisc; aniina; dérivation.
rection : LOC; LON; SCO; ACC; DAT; ABL;
ADA; ADA; ALO; AAB; DAB
dérivation : AGE; PPA; PPR; DVB; DAJ.
contdisc : CTN; DCT; CIN.
aniina : ANI; INA; ANN.
négativité : NEG; NNG; NGN.

ponctuation : PFOR, PFAI, PAUT.
PFOR(ponctuation forte: /./,/!/,/?/),
PFAI(ponctuation faible: /,/,/;/,/:/),
PAUT(autres ponctuations: /“/,/”/,/(/,/)/,...

Dans la phase de reconnaissance, seuls les formes sont concernées, ce qui justifie la simplicité des définitions. L'interprétation viendra en aval par imposition de la grammaire de réécriture du SN.
Dans l'analyse du texte, la méta-grammaire intervient dans le programme :
chaque forme du texte en entrée est associée un ensemble de solutions morphologiques, y compris les mots inconnus. Ces informations seront envoyées à l'étape suivante, l'analyse syntaxique.

Le texte est ainsi traité par les procédures qui suivent :
- le prétraitement de chaînes du texte,
- l'éclatement de toute chaîne isolée avec association d'information morphologique,
- le traitement général des apostrophes,
- une analyse d'une forme inconnue portera la marque de INDEF.

Puis, l'opération d'analyse flexionnelle qui consiste fondamentalement à ramener un fléchie à sa forme canonique. Le problème ainsi posé a permis de détecter trois types de flexions qui forment cinq cas par combinaisons : le pluriel, le féminin, le féminin pluriel, le genre non marqué et son pluriel.

III.4.5.3.4- La phase syntaxique

♣ Le formalisme des grammaires affixes

Dans la catégorie de **grammaires à deux niveaux**, nous avons les grammaires affixes, c'est-à-dire les grammaires qui utilisent une grammaire d'affixes appelée *méta-grammaire,* et une *hyper-grammaire*.
Autrement dit, le terme Grammaire affixes se rapporte aux deux grammaires hors-contextes d'une grammaire à deux niveaux, tandis que grammaire d'affixes ne fait allusion qu'à la méta-grammaire.
Pour avoir une idée sur les propriétés de grammaires affixes et tout particulièrement la grammaire EAG (Extended Affix Grammar) de D.A. Watt (1975), qui est issue du modèle W-Grammaires de V. Wijngaarden (1969) et du modèle Grammaire Affixale de C.H.A. Koster (1971) en les comparant avec les grammaires à deux niveaux :

Une EAG est définie comme un 10-uplet : EAG = $(V_N, V_T, A_N, A_T, e, R, B, D, S, P)$

V_N : un ensemble d'hyper non-terminaux,

V_T : un ensemble d'hyper terminaux,

A_N : un ensemble d'affixes non-terminaux,

A_T : un ensemble d'affixes terminaux,

$e \in V_N$: le symbole de départ (racine),

R: ensemble de règles d'affixes
B: ensemble de variables d'affixes,
D: une représentation $R \rightarrow A_N$,
S: l'ensemble de contrôle,
P: ensemble de hyper-règles.

Le premier niveau, généralement non-disjoint, comprend l'ensemble de la méta-grammaire hors-contexte : $G.a = \{ (A_N, A_T, a, R) / \forall a \in A_N \}$

Le deuxième niveau contient ce qu'il appelle la grammaire sous-jacente (UG: underlying grammar) : $UG = (V_N, V_T, e, P')$

où : P' est dérivé de P par omission des positions d'affixes,
P' est un ensemble de règles de production hors-contexte.
P : des hyper-règles qui contiennent la structure de base du langage.

♣ L'écriture des règles dans les grammaires affixes

Dans la grammaire d'affixes, la définition des types:

MS :: masc sing; nmgenre sing
FS :: fem sing; nmgenre sing
FP :: fem plu; nmgenre plu
MP :: masc plu; nmgenre plu

Dans la grammaire des notions, l'utilisation des affixes :

N"→ D' N' : D'(MS) N'(MS)
D'(FS) N' (FS)
D'(MP) N'(MP); etc.

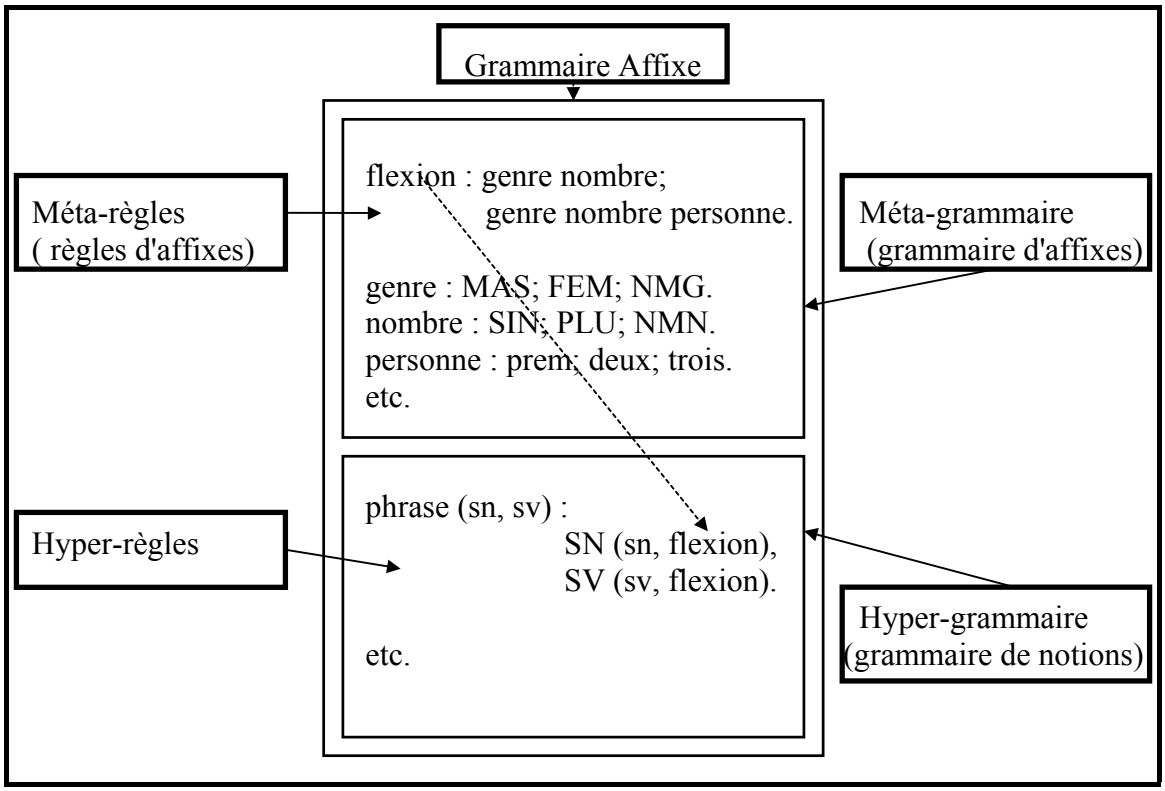


Fig.III.4.5.3.4. Formalisme des grammaires affixes.

Le prototype d'analyse syntaxique THAÏS (*Fig.III.4.5.3.4.*) est avant tout un outil qui pourra montrer d'autres perspectives d'analyse et faire évoluer l'état actuel des connaissances sur la grammaire de SN.

Certaines règles de la grammaire d'origine ont été légèrement modifiées pour deux raisons :

- minimiser les problèmes dus à l'absence d'un vrai lexique,
- tenter de corriger certaines imperfections de la grammaire du SN.

L'exemple de syntagme prépositionnel illustre que dans la grammaire du SN, il est reconnu par les règles suivantes:

$$N'' \rightarrow D' + N'$$
$$N' \rightarrow N + SP_n$$
$$SP_n \rightarrow SP^*$$
$$SP \rightarrow P + N''$$

Ces règles expriment une dépendance entre le SP et le SN (N"), c'est-à-dire le SP sera toujours rattaché au N : le centre du syntagme qui est tout de suite à sa gauche. Cette appartenance ne peut être prédite par les seules règles de réécriture mais que cela dépend des capacités rectionnelles des verbes et des nominaux (*cf.* [METZGER, 88], chap.X, p.187-190).

III.4.5.4- Conclusion

Transcrite dans le langage STARLET, la grammaire du syntagme nominal a été appliquée sur un corpus de plusieurs centaines de dépêches de l'AFP.

Les résultats obtenus montrent de façon significative la capacité offerte par les grammaires à deux niveaux de formaliser les phénomènes linguistiques.

Les avantages tirées de ces possibilités se manifestent, dans un premier temps, par la possibilité d'effectuer un contrôle plus strict sur l'application des règles d'analyse, et dans un deuxième temps, par les analyses qui peuvent être réalisées en absence du lexique.

III.5- Conclusion

Nous pensons qu'il est possible, dès à présent, de rendre compte à travers cette étude des présupposés théoriques antérieurs sur le statut du descripteur cherché pour indexer un document et accomplir un processus de recherche d'information. Ces présupposés théoriques sont fondés sur la détermination du processus discursif par ses conditions de production et le refus de la notion idéologique de *création infinis,* malgré sa mécanique de production qui se prête à ce genre d'extension.

Notre étude se focalise sur le discours à travers sa matérialité textuelle et s'oriente vers les unités organiques qui le compose sur les niveaux intensionnel et extensionnel. Nous ne prétendons pas fournir ici davantage d'orientations théoriques d'ordre linguistique, mais une esquisse de ce processus théoriquement fondé. Nous laissons au linguiste un grand nombre de décisions que nous ne pouvons prendre à sa place et nous retenons ses recommandations qui conditionnent le fondement théorique de son approche (cas de l'approche SYDO sur le statut du descripteur vs. le syntagme nominal). Les aspects pratiques conditionnent la deuxième phase de cette analyse et constituent à cet égard le véritable enjeu.

En résumé, ce processus exige un long travail au cours duquel des points de vues et des orientations pratiques seront avancés. L'essentiel était pour nous de spécifier ici le *requisit linguistiques* indispensables à l'analyse.

En d'autres termes, nous avons ici présenté le procédé de cette analyse à travers une approche théorique de la langue. Ce procédé repose sur des pré-supposés théoriques qui exigent précisément d'êtres explicites et ouvertes aux critiques. L'aspect critique nous l'avons exposé en partie à travers l'étude de quelques systèmes d'indexation.

CHAPITRE IV :

Extraction des connaissances dans un processus d'analyse et d'indexation des contenus : de l'Audiovisuel vers le Multimédia

IV. Introduction

L'audiovisuel est une forme d'expression comme l'écrit pour la connaissance, pour le savoir, pour la réflexion et la culture. Son mode d'expression est différent de celui de l'écrit. Son « alphabet » (ou ces primitives) se compose d'images animées et/ou de sons se déroulant de manière linéaire.

Contrairement au texte, l'image ne prescrit pas d'une manière directe une signification, mais elle peut s'interpréter en faisant appel aux conditions de sa conception et de ses lectures. Il est évident qu'un document, auquel il ne serait pas attaché des indications permettant sa lecture dans sa complétude (analyse, indexation et mémorisation), ne saurait franchir le cercle de sa réutilisation ou simplement de sa consultation [MANIEZ, 93].

L'existence physique d'un document et sa pérennité devrait disposer des informations pour son intégrité, pour son importance et pour la faculté de le retrouver et de le consulter [GUIMIER, 93], [BACHIMONT, 99a]. Dans un tel environnement, une synthèse professionnelle sur l'analyse de contenu documentaire s'avère nécessaire. Les technologies permettant la manipulation et l'extraction automatique des connaissances seront amenées à jouer un rôle essentiel dans la société de l'information.

IV.1- Influence de l'outil sur le traitement documentaire à l'I.N.A.

Les méthodes d'analyse des programmes audiovisuels [SIDHOM, 99] se sont structurées et développées au fur et à mesure de l'essor des moyens technologiques. Il apparaît évident qu'elles vont engendrer de nouvelles approches d'exploitation, de complétion et de restitution des connaissances (le contenu des programmes).

Depuis la création de la radio, de la télévision et des infrastructures du dépôt légal de l'audiovisuel, les recherches sur l'information et la communication par l'audiovisuel ont profondément évolué. La structuration d'une nouvelle communauté scientifique dans ce domaine a favorisé l'élargissement des études sur la télévision et la radio.

Quels savoirs issus de la recherche ayant comme sources la radio ou la télévision, et de quelles manières ?
Avant de répondre à cette question, nous étalons dans ce qui suit un bref historique des évolutions qui ont affecté ces sources de connaissances (en France) et les problématiques liées à l'indexation de l'audiovisuel [BACHIMONT, 2000].

IV.1.1- Historique et Evolution : 1949-1975

Les cinémathécaires, de façon régulière à partir de 1958, notent sur des cahiers, en direct, les contenus des émissions diffusées sur la 1ère chaîne, la 2ème puis la 3ème chaîne de la télévision française. Ils annotent les conducteurs, documents établis par le rédacteur en chef lors de la conférence de rédaction, décrivant le plan et la composition du journal télévisé. Ces conducteurs, source unique donnant le contenu, sont conservés et classés par ordre chronologique.
Le contenu de chaque émission (magazine, retransmission, etc.) fait l'objet d'une fiche dont le descriptif noté lors de la prise d'antenne est dactylographié. Ces fiches sont dupliquées puis classées par ordre chronologique et thématique.

En 1970, l'ORTF (Office national de la Radio et Télévision Française) charge le cabinet d'organisation SEMA (Documentation et gestion) de mettre au point une méthode d'analyse afin de gagner du temps sur la préparation de l'indexation des films d'actualités [LUSTIÈRE, 99]. L'analyse plan par plan, appelée « analyse chronologique », se révèle indispensable pour situer le sujet dans son contexte et faciliter la réponse à la demande (*Fig. IV.1.1.*).
La grille d'analyse chronologique mise au point à cette époque est utilisée jusqu'à présent à l'INA.
On peut y relever des informations élaborées (niveaux d'analyse) avec le concours des cinémathécaires sur un document donné:

- la position des plans et la description plan par plan,
- les mouvements de caméra,
- les personnes identifiées et les lieux de l'action,
- la distinction entre image et son.

Nombre de fiches : 17550
NOTICES CREEES A PARTIR DE FICHES RECONSTITUEES (Base conducteurs PRELUDA)

Id notice	: CAF97515400
Type notice (code)	: 31
Titre propre	: LA CRISE MINISTERIELLE
Titre collection (aff)	: JT 20H
Correspondant de chaîne	: STG
Auteurs (aff)	: JOU, DEBOUZY ROGER
Producteurs (aff)	: PRD, PARIS : RADIODIFFUSION TELEVISION FRANCAISE (RTF), 1950
Nature de production (aff)	: PRODUCTION PROPRE
Diffusion (aff)	: **05.07.1950 (D), 1 (ORTF)**
Durée (aff)	: 00H 03MIN 35SEC Environ
Résumé	: Après l'échec de Monsieur QUEUILLE, le Président de la République a repris des consultations - Pas de fiche dans le fichier chrono.
Notes	**: EX : N° TF0906I - Fiche reconstituée à partir du conducteur (PRELUDA)**
Anciens supports	: - No CINEMATHEQUE : FITF18024

Fig. IV.1.1. Exemple de Notice documentaire (source : INA-actualités).

IV.1.2- Historique et Evolution : 1975-1985

Le système IMAGO v.1. : Index des Média Audiovisuels Gérés par Ordinateur : ce système a pour but de retrouver une séquence de film dans la « mine » des archives audiovisuelles de la

télévision. Les données traitées par le système en font sa grande originalité. Il s'agit de documents filmiques : de séquences sonores et de vues animées.

Les cinémathécaires, devenus analystes de documentation à l'issue d'une formation sur le nouveau système (IMAGO), analysent en direct, séquence par séquence, sur des feuilles d'analyse chronologique les émissions diffusées sur TF1, Antenne 2 et France 3. Les différents niveaux du document sont codifiés. L'utilisation d'un thésaurus pour disposer d'une liste de mots à employer facilite la recherche d'une séquence d'images (*Fig. IV.1.2.*). Le thésaurus est composé de quatre langages documentaires :

- les noms communs,
- les noms propres géographiques,
- les noms propres de personnes morales,
- les noms propres de personnes physiques.

Les données traitées donnent lieu à trois types d'index édités sur microfiches :

- l'index analytique : descripteurs décrivant le contenu,
- l'index des titres : titres des séries, des émissions et des sujets,
- l'index générique : les noms des génériques des séries, des émissions et des sujets.

Pour aider à l'indexation et à la recherche, deux éditions du langage de description sont effectuées sous les noms :

- liste alphabétique du thésaurus, et
- liste méthodique du thésaurus.

Nombre de notices : 79043
Champs documentaires de bordereaux de saisie formatés

Titre propre	: VALERY GISCARD D'ESTAING-HELMUT SCHMIDT
Titre collection (aff)	: IT1 13H
Auteurs (aff)	: JOU, SAINT PAUL GERARD
Descripteurs principaux (aff)	: ALLEMAGNE RF ; HAMBOURG (OFF) ; GISCARD D'ESTAING VALERY ; SCHMIDT HELMUT
Producteurs (aff)	: PRD, PARIS : TELEVISION FRANCAISE 1 (TF1), 1978
Diffusion (aff)	: **24.06.1978, 1 (TF1)**
Forme	: JOURNAL TELEVISE
Résumé	: REUNION DE TRAVAIL ENTRE VGE ET SCHMIDT A HAMBOURG EN VUE D'ETABLIR UN TEXTE COMMUN SUR LA COOPERATION MONETAIRE EUROPEENNE...SAINT PAUL FAIT LE BILAN DE CET ENTRETIEN,VGE SORTANT DE VOITURE(DE NUIT), REJOINT PAR SCHMIDT.INTW VGE SUR LA TENEUR DE LEURS REFLEXIONS...
Anciens supports	: VICA U-MATIC,COULEUR : ACT-TF1 AK780624

Fig. IV.1.2. Exemple de Notice documentaire (source : INA-actualités).

IV. 1.3- Historique et Evolution : 1985-1997

Le système IMAGO v.2. : le système IMAGO évolue au mode conversationnel et l'interrogation est gérée avec le progiciel Mistral.
Les documentalistes continuent à analyser en direct, mais les données d'identification des émissions sont saisies par les personnes chargées de la collecte ou par les correspondants de chaîne. La base de données des archives TV (BDDTV) a pour objectifs de :

- gérer les descriptions documentaires des émissions TV,
- permettre l'interrogation en mode conversationnel de ces descriptions.

La base BDDTV constitue l'un des modules du système IMAGO v.2. Elle est intégrée aux modules suivants :

- module de collecte de données qui l'alimente, et
- module de gestion des stocks qui gère les supports relatifs aux émissions.

La base contient :

- l'ensemble des données de IMAGO v.1, soit les émissions nationales du 06/01/1975 au 25/08/1985, et leur reformatage selon le nouveau format normalisé de description des documents archives ;
- les descriptions des nouvelles émissions nationales diffusées à partir du 25/08/1985.

Le nouveau format est un compromis de toute une série de contraintes (de compatibilité avec la norme française, de complexité de la structure des émissions, de besoin en matière de recherche documentaire, etc.) et prévoit deux fonctionnalités (*Fig. IV.1.3.*) :

- le catalogage à deux niveaux : l'émission dans son ensemble ou le sujet de l'émission composite,
- l'indexation à deux niveaux : le thème général de l'émission ou l'événement auquel se rapporte le sujet (les champs résumé, descripteurs principaux, séquences, et descripteurs secondaires).

Interrogation sur base Imago-2, logiciel Mistral.
Saisie des champs textes en caractères riches sur terminaux d'ordinateur.

Titre propre	: NOUVELLE CALEDONIE : DIVERGENCES
Titre collection (aff)	: JA2 20H
Lien rediffusion (VA)	: CAB87005944 : RD JA2 DERNIERE.
Auteurs (aff)	: JOU, MERCUROL JEAN MICHEL
Descripteurs principaux (aff)	: TOM ; NOUVELLE CALEDONIE ; INDEPENDANCE ; REFERENDUM (OFF) ; AUTODETERMINATION ; LOI (PROJET) ; PONS BERNARD ; POLITIQUE INTERIEURE (COHABITATION) ; CONSEIL DES MINISTRES (OFF) ; MITTERRAND FRANCOIS (OFF)
Producteurs (aff)	: PRD, PARIS : ANTENNE 2 (A2), 1987 ; PRD, PARIS : RADIO FRANCE OUTRE MER (RFO), 1987
Diffusion (aff)	: **18.02.1987, 20H 01MIN 00SEC, 2 (A2)**
Durée (aff)	: 00H 01MIN 27SEC
Forme	: JOURNAL TELEVISE
Résumé	: F Mitterrand exprime son désaccord avec le gouvernement au sujet du projet de loi électorale fixant les conditions d'organisation du prochain référendum d'autodétermination en Nouvelle-Calédonie. - Mélanésien à cheval, case, enfants, piétons dans rue de Nouméa. Bernard Pons en visite en Nouvelle-Calédonie, Européens agitant des drapeaux tricolores, gens assistant à un meeting du RPCR. Groupe de Mélanésiens, fermier blanc à cheval et troupeau de vaches, ferme, palmeraie, femme tressant des feuilles de palmier pour en faire un panier.

Fig. IV.1.3. Exemple de Notice documentaire (source : INA-actualités).

IV.1.4- Historique et Evolution : 1997-1999

Le système BASIS Plus : la base d'IMAGO v.2 est reprise sous BASIS Plus. C'est un système de Gestion de Base de Données qui associe le système relationnel à la gestion documentaire.

On trouve deux types de notices documentaires :

- la notice d'émission : une notice documentaire d'une émission isolée, d'un sujet de journal télévisé ou d'un reportage traité comme un document,
- la notice d'ensemble : une notice documentaire de présentation et d'historique d'un ensemble d'émissions.

On trouve également différentes natures de champs :

- les champs réels : les données des notices (reprise des données Mistral et saisie avec PRODOC),
- les champs libellés : les libellés associés aux codes stockés dans les champs réels,
- les champs d'affichage : les champs virtuels réservés à l'affichage des valeurs calculées des champs réels (application BASINA),
- les index multi-champs : les champs réservés à la recherche par combinaison ou concaténation des valeurs issues des champs réels.

IV.1.5- vers le numérique dès le 3ème millénaire

Le passage au numérique est dans l'immédiat du perfectionnement des outils, nous citerons en exemples :

- Outils en usage ou en test à l'Inathèque : Vidéosribe, Médiascope
- Outils en usage à la vidéothèque de production : AGPE
- Outils en tests à la direction de l'Innovation : Eurodelphes

Ainsi la consultation du document se fera immédiatement après l'interrogation et la lecture des résumés dans la notice (*Fig. IV.1.5a. et Fig. IV.1.5b.*), voire même la consultation de la séquence cherchée.

Titre propre	**:** PLANETOLOGUE MARS
Diffusion (aff)	**:** 19.07.1997 (D), 20H 14MIN 55SEC, 2 (A2)
Durée (aff)	**:** 00H 01MIN 58SEC
Résumé	**:** Reportage sur Nathalie CABROL, planétologue française de la NASA, chargée de cartographier la planète MARS.
Séquences	**:** - DP Nathalie CABROL dans son laboratoire de recherches de la NASA à Pasadena étudiant des images de Mars en compagnie de ses collaborateurs.
	- ITW Nathalie CABROL à propose de son travail : "C'est une extension de la pensée humaine".
	- Nathalie CABROL devant une maquette du robot Sojourner / Suite ITW sur le fonctionnement du robot.
	- Nathalie CABROL dans son bureau devant un plan de Mars / VG dans son laboratoire avec d'autres chercheurs.

Fig. IV.1.5a. Notice documentaire de source : INA-actualités.

Titre propre	Que serions nous sans nos miroirs
Numéro	2510.001
Numéro DL	DL T 19950113 ART 005
Type de description	Emission simple
Statut de diffusion	Première diffusion
Société de programmes	ARTE
Canal	Réseau 5
Chaîne de diffusion	ARTE
Date de diffusion	13.01.1995
Jour	Vendredi
Heure de diffusion	22:22:35
Heure de fin de diffusion	23:14:11
Extension géographique	Multinational
Durée	00:51:36
Genre	Sciences
Forme	Documentaire
Auteurs	REA,Dars Jean François; REA,Papillault Anne; MUS,Artemyen Edouard
Générique	PAR,Carrière Jean Claude; PAR,Folon Jean Michel; PAR,Cyrulnik Boris
Producteurs	Producteur, Paris : Centre National De La Recherche Scientifique, 1995
Nature de production	Achat de droits de diffusion
Chapeau	Les miroirs sont le reflet de notre société. Psychologues et physiciens se penchent sur leurs multiples facettes.
Résumé	Au cours des premières minutes de reportage, on nous montre des collections de miroirs en bronze grecs, chinois, étrusques... Après avoir vu comment les miroirs étaient fabriqués industriellement, nous voyons de quelle manière on les travaille artisanalement. S'en suit une longue partie -durant laquelle de nombreux spécialistes interviennent- consacrée aux utilisations diverses du miroir à travers les âges : utilisation dans la fabrication des télescopes, de la lanterne magique, ancêtre du cinéma... L'accent est également mis sur les aspects psychologiques liés à la perception de soi.
Résumé producteur	On ne réfléchit jamais assez aux miroirs, c'est bien connu. Une équipe de réalisation du C.N.R.S. s'est donc penchée, pour le compte de LA SEPT/ARTE, sur ces étranges surfaces à la fois aveugles et brillantes : on trouve de tout dans les miroirs et il y a des miroirs partout.
Notes	Ce documentaire a obtenu le Prix Image-Unesco au 10ème Festival International de l'Emission Scientifique de Télévision en 1993.
Descripteurs	Foucault Léon; astronomie; télescope; miroir; trucage
Doc. d'accompagnement	Dossier de presse;Droits d'auteur
Matériel	BETA SP : 1 élément, Parallèle antenne, Couleur, MONO, Définition : 625 lignes, Format : 1/2 pouce, Signal : Analogique, Standard couleur : SECAM, Procédé : Béta
Titre matériel	Que serions nous sans nos miroirs
Couleur	Indéterminé
Repérage	22:21:00:00
Audience globale	0.60
Part de marché	2.40
Audience + 15 ans	0.70
Part de marché +15 ans	2.50
Audience homme	0.80
Audience femme	0.60
Type de traitement	Indexation de contenu
Version	Version Longue
Thèque	DL

Fig. IV.1.5b. Notice documentaire de source : INAthèque.

IV.2- Guide actuel de l'indexation audiovisuelle à l'I.N.A.

La grille d'analyse chronologique mise au point, à l'époque de l'ORTF et le cabinet SEMA en 1970 [DEGEZ, 70], est utilisée jusqu'à présent à l'INA. Les informations qu'on peut relever dans cette grille (ou niveaux d'analyse) sont :

- la position des plans et la description plan par plan,
- les mouvements de caméra,
- les personnes identifiées et les lieux de l'action,
- la distinction entre image et son.

On distingue plusieurs grilles d'analyse selon la nature de production tout en préservant le même principe de l'élaboration de la grille initiale, nous citons :

- PRODUCTION PROPRE (CODE 01)
- ACHATS DE DROITS (CODE 03)
- MIXTE (production propre + EVN ou achat ou archives) (CODE 04)
- EVN(CODE 05)
- ARCHIVES (CODE 10),

etc.

A- Construction des descripteurs par type de production

NATURE DE PRODUCTION	Descripteurs (DE)	Descripteurs Secondaires (DES)
PRODUCTION PROPRE **CODE 01**	Descripteurs thématiques et géographiques, personnes morales, personnes physiques évoquées.	Descripteurs séquences et images. Nom des villes si images réutilisables. Nom des personnes si visibles à l'image.
ACHATS DE DROITS **CODE 03**	Descripteurs thématiques et géographiques, personnes morales, personnes physiques constituant l'événement.	Pas de DES
EVN **CODE 05**	Descripteurs thématiques et géographiques, personnes morales, personnes physiques constituant l'événement.	Pas de DES
MIXTE (production propre + EVN ou achat ou archives) **CODE 04**	Descripteurs thématiques et géographiques, personnes morales, personnes physiques évoquées.	Descripteurs séquences et images. Nom des villes si images réutilisables. Nom des personnes si visibles à l'image.
MIXTE (chaînes privées, étrangères ou agences diverses) **CODE 04**	Descripteurs thématiques et géographiques, personnes morales, personnes physiques évoquées.	Pas de DES
ARCHIVES **CODE 10**	Descripteurs thématiques et/ou géographiques. Nom des personnes morales ou des personnes physiques en cas de rétro.	Pas de DES
etc.	…	…

B- Construction des résumés par type de production

NATURE DE PRODUCTION	Résumé court /chapeau (RES)	Résumé élaboré /séquence (SEQ)
PRODUCTION PROPRE **CODE 01**	Chapeau précisant la forme du sujet et situant l'événement dans le temps et dans l'espace	Description par séquence du sujet.

ACHATS DE DROITS **CODE 03**	Chapeau précisant la forme du sujet et situant l'événement dans le temps et dans l'espace	Pas de SEQ
EVN **CODE 05**	Chapeau précisant la forme du sujet et situant l'événement dans le temps et dans l'espace	Pas de SEQ
MIXTE **(production propre + EVN ou achat ou archives)** **CODE 04**	Chapeau précisant la forme du sujet et situant l'événement dans le temps et dans l'espace	Description par séquences du sujet, avec mention, si possible, des dates et de l'origine des plans d'archives. (INA, achats ou EVN)
MIXTE **(chaînes privées, étrangères ou agences diverses)** **CODE 04**	Chapeau précisant la forme du sujet et situant l'événement dans le temps et dans l'espace	Pas de SEQ
ARCHIVES **CODE 10**	Chapeau précisant la forme du sujet et situant l'évènement dans le temps et dans l'espace	Description par séquences du sujet, avec mention, si possible, des dates et de l'origine des plans d'archives. (INA, achats ou EVN).
etc.	…	…

Dans ce qui suit, nous reprenons quelques exemples de grilles d'analyses selon la nature de production avec illustration sur une notice produite à l'INA-actualités [LUSTIÈRE, 99].

IV.2.1- Grille d'Analyse : PRODUCTION PROPRE

NATURE DE PRODUCTION	DE	DES	RES	SEQ
PRODUCTION PROPRE CODE 01	Descripteurs thématiques et géographiques, personnes morales, personnes physiques évoquées.	Descripteurs séquences et images. Nom des villes si images réutilisables. Nom des personnes si visibles à l'image.	Chapeau précisant la forme du sujet et situant l'événement dans le temps et dans l'espace	Description par séquence du sujet.

Exemple 1 :

Titre propre :	ECLIPSE : CONSEILS SECURITE ROUTIERE
Titre collection :	19/20
Descripteurs principaux :	FRANCE ; POLITIQUE INTERIEURE ; ECLIPSE ; SOLEIL ; PREVENTION ; CAMPAGNE D'INFORMATION ; SECURITE
Descripteurs secondaires:	LUNETTES ; HOMME POLITIQUE ; FEMME ; MINISTRE ; DEMESSINE MICHELLE ; AUBRY MARTINE ; BUFFET MARIE GEORGE ; GAYSSOT JEAN CLAUDE ; LIEBART BERNARD ; ANIMATION ; CIRCULATION ROUTIERE
Producteurs (aff) :	PRD, PARIS : FRANCE 3 (F3), 1999
Nature de production (aff):	PRODUCTION PROPRE
Résumé :	Présentation des mesures de sécurité recommandées, notamment en matière de circulation routière, lors de l'éclipse solaire du 11 août prochain.
Séquences :	- Sur le perron de l'Observatoire de PARIS, quatre MIN (Michèle DEMESSINE, Martine AUBRY, Marie-George BUFFET et Jean-Claude GAYSSOT), posent avec des lunettes "spéciales éclipse" / conférence de presse. - SYNTHE CARTE montrant l'amplitude de l'éclipse. - DP circulation sur autoroute.

	- ITW de Bernard LIEBART (vice-PDT de la Fédération nationale des transporteurs routiers - PDG des Transports Liebart), mécontent du fait que le gouvernement leur demande de ne pas rouler pendant 5 heures le 11 août prochain : "Je vais peut-être envoyer la facture à Mr Jospin". - ITW d'un chauffeur routier. - Circulation routière. - GP Jean-Claude GAYSSOT (MIN de l'Equipement, des Transports et du Logement). - SYNTHE ANIMATION éclipse / GP lunettes.

Exemple 2 :

Titre propre :	ST ETIENNE/FEU D'ARTIFICE MORTEL
Titre collection :	JA2 20H
Descripteurs principaux :	FRANCE ; LOIRE ; SAINT JUST SAINT RAMBERT ; FETE NATIONALE (14 JUILLET) ; FEU D'ARTIFICE ; INCIDENT ; DECES ; BLESSE ; ENQUETE POLICIERE
Descripteurs secondaires :	GENDARME ; ARVISET BRUNO ; INTERVIEW ; TEMOIGNAGE ; CHOSSY JEAN FRANCOIS
Producteurs (aff) :	PRD, PARIS : FRANCE 2 (F2), 1999
Nature de production (aff):	**PRODUCTION PROPRE**
RESUME:	Une femme a été tuée et sept personnes blessées par une fusée tirée lors du feu d'artifice du 14 juillet de Saint-Just-Saint-Rambert (Loire).
SEQUENCES:	- Saint-Just-Saint-Rambert (Loire): DP policiers et badauds à l'endroit où a été tiré le feu d'artifice, sur la rive de la Loire - Témoignage d'une spectatrice racontant que des morceaux sont tombés sur la foule - ITW Capitaine Bruno Arviset, cdt de la compagnie de gendarmerie de Montbrison (Loire), expliquant qu'un tréteau était sans doute mal fixé - Gendarmes ramassant des débris et prenant des photos - ITW du député maire, Jean-François Chossy, expliquant que c'est la même société qui tire le feu d'artifice depuis des années

IV.2.2- Grille d'Analyse : ACHATS DE DROITS

NATURE DE PRODUCTION	DE	DES	RES	SEQ
ACHATS DE DROITS CODE 03	Descripteurs thématiques et géographiques, personnes morales, personnes physiques constituant l'événement.	Pas de DES	Chapeau précisant la forme du sujet et situant l'événement dans le temps et dans l'espace	Pas de SEQ

Exemple :

Titre propre :	TOUT IMAGES : APPONTAGE CHARLES DE GAULLE
Titre collection :	SOIR 3
Descripteurs principaux :	**FRANCE ; PORTE AVIONS (CHARLES-DE-GAULLE) ; AVION DE**

	CHASSE (RAFALE) ; DECOLLAGE
Producteurs (aff) :	PRD, IVRY : ETABLISSEMENT PUBLIC CINEMATOGR. DES ARMEES (ECPA), 1999
Nature de production (aff):	**ACHAT DE DROITS DE DIFFUSION**
RESUME:	LE « CHARLES-DE-GAULLE » ACCUEILLE SON PREMIER RAFALE. Pour la première fois, un avion RAFALE a pu se poser, puis redécoller du nouveau porte-avions nucléaire français, croisant au large de Brest. L'appontage du Rafale a eu lieu mardi 6 juillet 1999 et son catapultage mercredi. Le CHARLES-DE-GAULLE doit subir encore une série d'essais. Après une dernière croisière d'endurance, il devrait être déclaré opérationnel fin 2000.

IV.2.3- Grille d'Analyse : EVN

NATURE DE PRODUCTION	DE	DES	RES	SEQ
EVN **CODE 05**	Descripteurs thématiques et géographiques, personnes morales, personnes physiques constituant l'événement.	Pas de DES	Chapeau précisant la forme du sujet et situant l'événement dans le temps et dans l'espace	Pas de SEQ

Exemple 1 :

Titre propre :	OFF KHROUCHTCHEV DEMANDE NATIONALITE USA
Titre collection :	MIDI 2
Descripteurs principaux:	ETATS UNIS ; KHROUCHTCHEV SERGUEI ; REFUGIE ; NATIONALITE Producteurs (aff) :PRD, LONDRES : ASSOCIATED PRESS TELEVISION NEWS (APTN), 1999
Nature de production :	**ECHANGES INTERNATIONAUX EVN**
Résumé :	Le fils de Nikita KHROUCHTCHEV, Sergueï, est devenu citoyen des Etats-Unis en prêtant serment à RHODE ISLAND.

Exemple 2 :

Titre propre :	OFF KOSOVO/MANIF
Titre collection :	JA2 DERNIERE
Descripteurs principaux :	REPUBLIQUE FEDERALE DE YOUGOSLAVIE ; KOSOVO ; MANIFESTATION ; ALBANAIS ; ARMEE (RUSSIE)
Producteurs (aff) **:**	PRD, ROYAUME UNI : REUTERS TELEVISION (RTV), 1999
Nature de production (aff):	**ECHANGES INTERNATIONAUX EVN**
Résumé :	Dans le sud-est de la province du Kosovo, zone sous contrôle des soldats allemands, des Albanais ont manifesté aujourd'hui pour faire part de leurs craintes de voir les soldats russes de la KFOR prendre le parti des Serbes restés dans la zone.

IV.2.4- Grille d'Analyse : MIXTE (production propre + EVN ou achat ou archives)

NATURE DE PRODUCTION	DE	DES	RES	SEQ
MIXTE (prod propre + EVN ou achat ou archives) **CODE 04**	Descripteurs thématiques et géographiques. personnes morales, personnes physiques. évoquées.	Descripteurs séquences et images. Nom des villes si images réutilisables. Nom des personnes si visibles à l'image.	Chapeau précisant la forme du sujet et situant l'événement dans le temps et dans l'espace	Description par séquences du sujet, avec mention, si possible, des dates et de l'origine des plans d'archives. (INA, achats ou EVN)

Exemple :

Titre propre :	INDEMNISATION VICTIMES ATTENTAT
Titre collection :	19/20
Descripteurs principaux:	RELATIONS DIPLOMATIQUES (FRANCE LIBYE) ; AVION (DC10) ; TERRORISME ; CATASTROPHE (AERIENNE) ; ATTENTAT ; COMPAGNIE AERIENNE (UTA) ; RETROSPECTIVE ; INDEMNISATION ; VICTIME
Descripteurs secondaires:	INTERVIEW ; RUDETZKI FRANCOISE
Producteurs (aff) :	PRD, PARIS : FRANCE 3 (F3), 1999 ; PRD, BRY SUR MARNE : INSTITUT NATIONAL DE L'AUDIOVISUEL (INA), 1999
Nature de production (aff):	**MIXTE**
RESUME :	FRANCE-LIBYE Attentat DC-10 d'UTA: l'indemnisation : La LIBYE a transféré en FRANCE une somme de plus de 200 millions de FF destinée à indemniser les familles des 170 victimes de l'attentat contre un DC-10 d'UTA en 1989 au Niger. Cette somme est conforme aux arrêts prononcés en mars dernier par la cour d'assises de Paris, qui avait aussi condamné par contumace six fonctionnaires libyens à la perpétuité. Si le gouvernement français s'est dit "satisfait" du versement vendredi, SOS-Attentats a jugé de son côté la somme insuffisante et a déploré l'impunité des accusés.
SEQUENCES :	- Archives INA Désert du Ténéré au Tchad: Débris du DC10 d'UTA. - Archives non siglées: DP Photo du épave du DC 10. - Archives non siglées: PANO Juge d'instruction Jean Louis BRUGUIERE. - Archives non siglées: BT photo noir et blanc d'Abdallah SENOUSSI. - Archives non siglées: PR Mohammed El Kadhafi. - Archives non siglées: DP Juge Jean louis BRUGUIERE. - Archives non siglées: DP de Kadhafi. - ITW Françoise RUDETZKI Présidente de l'association SOS attentat: "C'est absolument pas satisfaisant, c'est insultant ce que les victimes demandent c'est la justice. Que la sanction pénale prononcée contre les six hauts responsables libyens soit exécutée." - DP Cimetière avec un monument dédiés aux victimes de l'attentat.

IV.2.5- Grille d'Analyse : MIXTE (chaînes privées, étrangères ou agences diverses)

NATURE DE PRODUCTION	DE	DES	RES	SEQ
MIXTE (chaînes privées, étrangères ou agences diverses) **CODE 04**	Descripteurs thématiques et géographiques, personnes morales, personnes physiques. évoquées.	Pas de DES	Chapeau précisant la forme du sujet et situant l'événement dans le temps et dans l'espace	Pas de SEQ

Exemple :

Titre propre :	LES SEPT MISSIONS LUNE
Titre collection :	MIDI 2
Descripteurs principaux :	ETATS UNIS ; ESPACE ; ASTRONAUTIQUE ; LUNE ; NAVETTE SPATIALE (APOLLO) ; BILAN ; RETROSPECTIVE
Producteurs (aff) :	PRD, WASHINGTON : NATIONAL AERONAUTIC AND SPACE ADMINISTRATION (NASA), 1999
Nature de production (aff):	**MIXTE**
RESUME :	Les Américains ont effectué 7 missions sur la Lune dans les capsules "Apollo" pendant 2 ans à compter de la première du 20 juillet 69. Qu'en reste-t-il aujourd'hui ?

IV.2.6- Grille d'Analyse : ARCHIVES

NATURE DE PRODUCTION	DE	DES	RES	SEQ
ARCHIVES **CODE 10**	Descripteurs thématiques et/ou géographiques. Nom des personnes morales ou des personnes physiques en cas de rétro.	Pas de DES	Chapeau précisant la forme du sujet et situant l'évènement dans le temps et dans l'espace	Description par séquences du sujet, avec mention, si possible, des dates et de l'origine des plans d'archives.(INA, achats ou EVN).

Exemple :

Titre propre :	OFF OLIVIER GUICHARD/MISE EN EXAMEN
Titre collection :	MIDI 2
Descripteurs principaux :	**FRANCE ; JUSTICE ; HOMME POLITIQUE ; GUICHARD OLIVIER ; MISE EN EXAMEN**
Producteurs (aff) :	PRD, PARIS : FRANCE 2 (F2), 1999
Nature de production :	**DOCUMENTS ARCHIVES**
RESUME :	l'ancien baron du gaullisme, Olivier GUICHARD a été mis en examen pour notamment "abus de confiance et d'ingérence" . Les faits qui lui sont reprochés remontent à la période où il présidait le conseil régional des pays de Loire .
SEQUENCES :	- Arc. PM de O.GUICHARD sortant d'un bâtiment .

IV.3- Constitution d'un corpus de notices documentaires à l'I.N.A.

IV.3.1- Objectifs de l'étude d'un corpus (notices documentaires)

Une application multimédia est composée de données hétérogènes : textes, sons, graphiques, images fixes, animations, vidéos. Dans leur organisation, en vue d'une exploitation documentaire, les connaissances manipulables sont contenues dans les textes attachés à ces documents [BICHARD, 92] ou les parties composites du document (annotations).
La constitution d'un corpus de documents hétérogènes (audiovisuels) contenant des textes résumés a pour but, dans une première étape de notre travail de recherche, d'opérer des analyses et de relever des régularités pour des formulations syntaxiques. La nature même de

ces résumés n'est pas construite ad-hoc, mais selon des principes fondés et une gestion d'analyse du contenu convenablement construits [CLAVEL, 93]. Cette procédure situe le sujet du document et ses différentes parties dans leur contexte.

Les principales étapes du processus consistent en un recueil de notices documentaires incorporant des résumés de contenus chez des fournisseurs spécialisés dans ce domaine.
Notre collaboration scientifique avec des spécialistes de l'INA (Institut National de l'Audiovisuel en France), a révélé une expérience professionnelle et des acquis qui datent des années cinquante.
Actuellement, la formulation des résumés sur des contenus documentaires hétérogènes est construite, dans certains organismes spécialisés comme l'INA, selon des critères et des méthodes formelles acquises par l'expérience [BROWNE, 96]. Cela permet d'assurer une régularité et une constance des traitements réalisés par les documentalistes. Cette richesse documentaire, une fois construite et mise à l'exploitation selon des traits attachés au contenu (capitalisation des sources d'information et de connaissances), pourra s'adapter aux diverses technologies d'exploitation et de diffusion [CHAMPENIER, 96], [PINON, 96], [MARET, 94].

IV.3.2- Caractéristiques des notices documentaires

IV.3.2.1- Notice d'un document audiovisuel (émissions TV) : Corpus INAthèque

N° Champ	Nom du Champ	Type	Exemples
1	Titre propre	Texte (titres)	1. Un lac venu de l'espace :... 2. [Le monde des hélicoptères : émission du 14 janvier 1995] 3. L'oeuvre scientifique de Pasteur = Das wissenschaftliche Werk Pasteurs
2	Titre collection	Texte (titres)	Le monde des hélicoptères ;
3	Titre programme	Texte (titres)	Les cinq continents
4	Numéro		842.001
5	Numéro DL		DL T 19950101 FR2 022
6	Type de description		Emission simple
7	Statut de diffusion		Diffusé
8	Société de programmes		France 2
9	Canal		Réseau 2
10	Chaîne de diffusion		France 2
11	Date de diffusion		01.01.1995
12	Jour		Dimanche
13	Heure de diffusion		27:55:41
14	Heure de fin de diffusion		28:24:11
15	Extension géographique		National
16	Durée		00:28:30
17	Genre		Sciences
18	Forme		Documentaire
19	Auteurs		REA,Larochelle André; SCE,Bouchard Michel; ATP,Pelletier Pierre
20	Générique		COM,France Ronald
21	Générique secondaire		PAR,Sobotta Siegdried; PAR,Vegter Nico
22	Producteurs		Producteur, Paris : France 3, 1994;Saint Ouen : Gédéon, 1994;Paris : ELF, 1994

23	Nature de production		Production propre
24	Chapeau	Texte libre	Ce documentaire retrace les travaux menés par une équipe de chercheurs dans le nouveau Québec, afin d'expliquer la présence d'un lac qui se serait formé suite à la chute d'une météorite.
25	Résumé	Texte libre	La chute d'une météorite venue de l'espace a créé un lac dans la Toundra du Nouveau Québec. Celui-ci mesure 2,7 km de diamètre, 267 mètres de profondeur et son cratère s'étend sur 3 km.
26	Séquences	Texte (titres)	-DP Mirage en looping. -GP tête du pilote dans le cockpit. -Vieux coucou et Mirage faisant loopings côte à côte. -Auscultation par un médecin de JM Denuel. -Denuel se préparant au décollage, décollage. -Parachute s'ouvrant à l'atterrissage.
27	Résumé producteur	Texte libre	On ne réfléchit jamais assez aux miroirs, c'est bien connu. Une équipe de réalisation du C.N.R.S. s'est donc penchée, pour le compte de LA SEPT/ARTE, sur ces étranges surfaces à la fois aveugles et brillantes : on trouve de tout dans les miroirs et il y a des miroirs partout.
28	Notes de titre		Dépôt des cendres de Pierre et Marie Curie au Panthéon.
29	Notes	Texte libre	La version d'une heure de "Un lac venu de l'espace" a remportée le Prix de la meilleure vulgarisation scientifique au Festival International du film scientifique de Palaiseau en 1990.
30	Descripteurs	Thésaurus INA	météorite; lac (Nouveau Québec); Québec; expédition (scientifique); chercheur
31	Descripteurs secondaires	Thésaurus INA	sciences humaines; enseignement
32	Doc. d'accompagnement		Fiche technique
33	Matériel		BETA SP : 1 élément, Parallèle antenne, Couleur, MONO, Définition : 625 lignes, Format : 1/2 pouce, Signal : Analogique, Standard couleur : SECAM, Procédé : Béta
34	Titre matériel		Un lac venu de l'espace : le cratère du Nouveau Québec
35	Couleur		Couleur
36	Repérage		04:10:00:00
37	Audience globale		0.10
38	Part de marché		25.00
39	Audience + 15 ans		0.20
40	Part de marché +15 ans		40.00
41	Audience homme		0.10
42	Audience femme		0.20
43	Type de traitement		Indexation de contenu
44	Version		Version Longue
45	Thèque		DL

IV.3.2.2- Notice document écrit : Corpus INAthèque

N° Champ	Nom du Champ	Type	Exemples
1	Analyste		PHR
2	Auteur(s)		MAUSSET (Martine)
3	Bibl/ Index		Index
4	Collation		p 63 - p 65
5	Cote CDU		62-527 AHR
6	Date parution		MARS 88
7	Descripteurs	Thésaurus INA	TECHNIQUE; NOUVELLE TECHNOLOGIE; DOMOTIQUE; FRANCE
8	Domaine		AVC
9	Editeur		Groupe de liaison pour l'action culturelle scientifique
10	Lieu d'édition		Paris, Liège
11	Langue		Francais
12	Mis à jour le		1995-12-15
13	Mis à jour par		SAUID
14	N° Inventaire		36636
15	N° Référence		2307
16	Résumé	Texte libre	L'habitat est-il à l'aube d'une nouvelle révolution ? quoi qu'il en soit,les fonctions de sécurité, maitrise de la consommation d'énergie, modification des modes de communication, sans oublier la télésurveillance, télécommande etc.
17	Revue		Videotex n 33
18	Saisi le		1988-03-21
19	Titre	Texte (titres)	Domotique : des maisons intelligentes & communicantes
20	Traduction		Claude Garrot

IV.3.2.3- Notice document audio (émission radio) : Corpus INAthèque

- Notice Information Sciences

N° Champ	Nom du Champ	Type	Exemples
1	Titre propre	Texte (titres)	Info Sciences
2	Titre collection	Texte (titres)	Le téléphone sonne
3	Titre programme		France Info Express du 07 février 1995
4	Numéro		17230.004
5	Numéro DL		DL R 19950216 FIT 12
6	Type de description		Sujet
7	Statut de diffusion		Diffusé
8	Société de programmes		Radio France
9	Canal		Modulation de fréquence réseau 6
10	Chaîne de diffusion		France Info
11	Date de diffusion		07.02.1995
12	Jour		Mardi
13	Heure de diffusion		00:51:00
14	Heure de fin de diffusion		24:53:00
15	Durée		24:02:00

16	Forme		Chronique
17	Public destinataire		Participation du public à distance
18	Générique		PRE,Monchicourt Marie Odile
19	Auteurs		PRO,Bedouet Alain; REA,Rosier Catherine
20	Date d'enregistrement		07.02.1995
21	Producteurs		Producteur, Paris : Radio France, 1995
22	Nature de production		Production propre
23	Chapeau	Texte libre	Marie Odile MONCHICOURT évoque la lutte contre les termites, véritable fléau de plusieurs pays africains. Elle parle des avancées de la recherche : création de pièges pour les termites ailées et introduction de fongicides destinés à erradiquer des champignons indispensables aux termites.
24	Résumé	Texte libre	1. ITW DU DOCTEUR AMALBERTI, CHERCHEUR AU MINISTERE DE LA DEFENSE, ALAIN GRAS PROF. A LA SORBONNE, MARC LABRUCHERIE COMMANDANT DE BORD A AIR FRANCE, JEAN-CLAUDE WANNER PILOTE MILITAIRE, HUGHES GENDRE PSDT DU SYNDICAT NATIONAL DES PILOTES DE LIGNE. AVEC LA VOIX DE SAINT-EXUPERY. L'IMPORTANCE 2. Ces cavernes, dont les murs sont recouverts de fresques préhistoriques, furent fréquentées par des visiteurs dès le 17ème siècle. Mais il fallut attendre 1906 pour que les gens prennent conscience de son importance historique, grâce à Félix GARRIGOU, préhistorien.
25	Descripteurs	Thésaurus INA	science; faune; termite
26	Matériel		CD Worm. MUSICAM 96, Mono. Parallèle antenne
27	Titre matériel		France Inter, 16 Février 1995, 12 h - 23 h
28	Repérage		00:51:00
29	Type de traitement		Indexation de contenu
30	Thèque		Dépôt Légal

- Notice Journal-Parlé Sciences

N° Champ	Nom du Champ	Type	Exemples
1	Titre propre	Texte (titres)	La somnolence au volant
2	Titre collection	Texte (titres)	Le téléphone sonne
3	Titre programme		Inter soir 19H00 du 02 juillet 1995
4	Numéro		40722.013
5	Numéro DL		DL R 19950702 FIT 12
6	Type de description		Sujet
7	Statut de diffusion		Première diffusion
8	Société de programmes		Radio France
9	Canal		Modulation de fréquence réseau 1

10	Chaîne de diffusion		France Inter
11	Date de diffusion		02.07.1995
12	Jour		Dimanche
13	Heure de diffusion		19:00:00
14	Heure de fin de diffusion		19:00:45
15	Durée		00:00:45
16	Forme		Journal parlé
17	Public destinataire		Participation du public à distance
18	Générique		PRE,Monchicourt Marie Odile
19	Auteurs		PRO,Bedouet Alain; REA,Rosier Catherine
20	Date d'enregistrement		07.02.1995
21	Producteurs		Producteur, Paris : Radio France, 1995
22	Nature de production		Production propre
23	Chapeau	Texte libre	Une étude de l'université de Boston tente de démontrer les rapports entre canicule et taux de natalité.
24	Résumé	Texte libre	La théorie, baptisée "+5 -6", prouverait que lorsque la température monte de 5 degrés, 9 mois plus tard la natalité baisse de 6%. Microtrottoir dans les rues de Paris pour savoir si la chaleur intervient sur la libido.
25	Descripteurs	Thésaurus INA	science; faune; termite
26	Matériel		CD Worm. MUSICAM 96, Mono. Parallèle antenne
27	Titre matériel		France Inter, 16 Février 1995, 12 h - 23 h
28	Repérage		00:51:00
29	Type de traitement		Indexation de contenu
30	Thèque		Dépôt Légal

IV.3.2.4- Notice document image

- Notice Image Fixe : Constitué sur Internet

N° Champ	Nom du Champ	Type	Exemples
1	* Numéro		1.
2	Titre	Texte (titres)	L'albertosaure
3	Titre Collection	Texte (titres)	Le musée des dinosaures
4	Résumé producteur	Texte libre	L'albertosaure a été découvert en Alberta, au Canada, d'où l'origine de son nom. Il appartient à la même famille que le tyrannosaure rex et, comme ce dernier, c'est un grand carnivore. À l'endroit où il a été trouvé, on a créé un Parc national de dinosaures. Plus de 350 squelettes de dinosaures ont été découverts à cet endroit.
5	Signification	Texte (titres)	Signification de son nom: lézard de l'Alberta.
6	Site Web		http://www.globetrotter.qc.ca/escale/dinos/musee01.htm
7	Adresse Web		http://www.globetrotter.qc.ca/escale/dinos/musee02.htm

- Notice Image Animée : Corpus INA-actualités

N° Champ	Nom du Champ	Type	Exemples
1	Id notice		CAB97119503
2	Titre collection	Texte (titres)	ESPACE.
3	Titre propre	Texte (titres)	PLANETOLOGUE MARS
4	Diffusion		20H 14MIN 55SEC, 2 (A2)
5	Durée		00H 01MIN 58SEC
6	Nature de production	Code	**DOCUMENTS ARCHIVES**
7	Résumé	libre	Reportage sur Nathalie CABROL, planétologue française de la NASA, chargée de cartographier la planète MARS.
8	Séquences	Texte libre	- DP Nathalie CABROL dans son laboratoire de recherches de la NASA à Pasadena étudiant des images de Mars en compagnie de ses collaborateurs. - ITW Nathalie CABROL à propose de son travail : "C'est une extension de la pensée humaine". - Etc.

IV.3.3- Principales caractéristiques des composantes d'une notice documentaire de l'audiovisuel

L'information inscrite dans les notices documentaires, cas de l'INA, est variée et complémentaire. Les données inscrites sont partagées sur plusieurs critères à savoir : l'acquisition, la production, l'usage et le contenu du document (pour le savoir et la connaissance, etc.).
Le dernier critère, le contenu du document, a fait l'objet de notre étude. Nous nous intéresserons aux champs contenant du texte libre sur l'analyse du contenu. Nous citerons, en particulier pour les notices de l'INAthèque, les champs suivants :

- titre propre : titre du document
- chapeau : résumé court et général sur le contenu
- résumé : résumé détaillé
- séquences : description thématique par séquence
- résumé producteur : résumé source par l'auteur ou par la société productrice

L'élaboration des champs résumés (chapeau, résumé, séquences) ont fait l'objet d'études scientifiques et d'expertises professionnelles. Pour l'INA, il s'agit des grilles d'analyse des documents de l'audiovisuels (*cf.* chapitre IV.1.).

L'indexation et la structuration documentaires mobilisent différentes dimensions pour lesquelles différents objets de ces dimensions sont mis en exergue [FGAIER, 92] [FLUHR, 92-97]. Le travail qui s'accomplit en ce domaine est un travail de fond dont les résultats progressent lentement : la représentation du contenu des documents non-textuels, leur indexation, et la recherche d'information sur des documents hétérogènes [DJEBAR, 89] [DESAI, 95] [DHENIN, 96] [GEDIMAGE, 96] semblent, par contraste, être l'objet d'une activité beaucoup plus intense et quasiment monopoliser les efforts de la recherche. Pour l'heure, il s'agit d'un édifice de recherche théorique interdisciplinaire.

IV.3.4- Savoirs théoriques et réalités professionnelles

La recherche théorique vise à la constitution d'un savoir. Dans chaque domaine théorique, il est question de définir, d'opérer et de retrouver des objets ou de nouveaux objets. C'est ainsi que le savoir se véhicule par le biais d'un support.

L'objet d'une recherche théorique est celui constitué et construit par cette recherche. Il se définit par les lois auxquelles il obéit : il est le corrélat des lois de son domaine.

En mathématiques, la construction d'un concept signifie la définition d'une méthode permettant de construire un objet.
En physique-mathématiques, l'objet physique se définit par les équations qui le régissent et qu'il obéit. Ces équations sont établies par la confrontation d'une multitude de mesures expérimentales et à travers sa médiation.
En linguistique, les objets théoriques de la langue n'ont pas d'existence empirique, mais des abstractions ou des idéalisations pour formuler des lois.

Dans le cadre d'une recherche théorique portant sur des objets issus d'une pratique effective, ces objets ne sont pas constitués par la recherche, mais par des activités humaines sociales, professionnelles, ou technologiques.
Dans le cadre de ces pratiques, on peut aborder un objet concret et rencontrer sa complexité. De ce point de vue, l'objet concret ou réel est véritablement interdisciplinaire. L'objet théorique ou le théorique reste encadré dans les limites des ses lois fondatrices. Dans cette perspective, les recherches définies à partir d'une pratique ou de métier sont importantes sur le plan de l'économie et de la gestion du savoir.

Il importe bien de comprendre qu'une recherche sur le multimédia [FLEXER, 91] et particulièrement sur l'audiovisuel qui hérite des savoirs théoriques [LESPINASSE, 2001] et confrontée à des réalités professionnelles du domaine, pourra constituer un édifice de recherche théorique interdisciplinaire [BACHIMONT, 99b].

- L'audiovisuel comme objet physique

C'est le cadre du traitement des images et des sons. Un objet audiovisuel est un support physique ayant une structure pour enregistrer une information audiovisuelle. Si on se place du côté des technologies numériques, cette approche renvoie aux primitives d'accès, d'enregistrement, de modification et de suppression aussi bien aux techniques de marquage et de filigranage des images et des sons. L'enjeu sur cet objet est de traiter et travailler sur le code de cette information physique : numérique ou analogique.

- L'audiovisuel comme objet analysable

Fait l'objet de l'analyse du flux d'information audiovisuelle. L'information inscrite sur le support physique est exploitable pour permettre une analyse et pour rendre compte des ses éléments qui la composent. Dans ce cas, c'est la décomposition du flux en images et en sons. Cette analyse permet de séparer les constituants et d'obtenir les propriétés physiques analysables du contenu audiovisuel : texture, couleur, formes etc.

- L'audiovisuel comme objet interprétable

Il est question de la structuration et de l'indexation documentaire. Les objets audiovisuels reçoivent différentes descriptions documentaires pour permettre l'extension de la dimension de leur utilisation. Pour accéder au contenu, l'analyse, l'indexation, l'annotation et la

structuration réalisés par les documentalistes permettent d'accéder partiellement ou totalement aux contenus audiovisuels selon différentes perspectives. Sur les plans méthodologique et technique, il s'agit de l'emploi des canaux de la terminologie, de la linguistique, de l'expertise documentaire permettant de décrire le contenu et d'intégrer les résultats pour une exploitation automatique du flux.

- L'audiovisuel comme objet de médiation

Il s'agit des outils de gestion et de consultation du flux audiovisuel. Le numérique banalise l'exploitation et la gestion du flux tant en mode de réception et de transmission, aussi bien en mode de consultation et d'appropriation du contenu du flux. Les nouvelles technologies numériques actuelles offrent certains outils sophistiqués de gestion et d'exploitation appropriés aux contenus audiovisuelles et d'autres qui nécessitent l'augmentation des performances. Il s'agit de mobiliser plusieurs techniques pour la profusion et la gestion de la complexité des informations audiovisuelles dont le but est de permettre à l'utilisateur l'appropriation ou l'intégration des contenus consultés pour une nouvelle valorisation.

- L'audiovisuel comme objet technique

L'étude et l'analyse des usages des médiations techniques de l'audiovisuel constituent des approches fondatrices pour la conception et l'amélioration des techniques et des nouveaux outils. En sachant pour qui, pourquoi et comment, que l'on peut fructifier et valoriser l'accès au savoir contenu dans l'audiovisuel. Plusieurs méthodologies consistent sur le plan technique à focaliser l'observation des usagers et les outils en phase d'interaction pour rendre analysable les scénarii d'interaction et pour dégager ainsi des critères et des anomalies pouvant adapter l'innovation à la demande.

- L'audiovisuel comme objet culturel

L'analyse socio-économique des programmes audiovisuels est importante. L'audiovisuel est un vaste système qui confirme la culture du savoir dans diverses disciplines à travers les contraintes sur l'élaboration, la diffusion, la transmission et la conservation des programmes. La notion de programme audiovisuel, point d'évaluation de l'offre et de son public, fournit un outil d'évaluation, un outil d'interprétation et d'aide à la décision dans le cadre des implications culturelle, économique et sociale. L'audiovisuel est un espace de savoirs [BACHIMONT, 92] et de l'émergence de techniques qui fournissent des considérations économiques et culturelles [BACHIMONT, 99c].

Remarques :

Les progrès en ce domaine, la représentation des documents, sont pourtant nécessaires car il ne fait pas de doute que les outils bureautiques, la gestion électronique des documents (GED) et les moteurs de recherche sur le Web vont focaliser et rénover pour une part importante des efforts de l'informatique documentaire. Les problèmes posés sont stimulants : documents multimédia, absence d'intermédiaire documentaliste, intégration des outils documentaires ou de recherche d'information autonomes [BACHIMONT, 99a].

IV.4- Etude statistique sur le corpus

Nous avons cherché à étudier la stabilité des descriptifs textuels (résumés), tout particulièrement ceux dans les notices INA (sources de INAthèque puis INAactualités), afin d'établir par une analyse statistique les composantes grammaticales et syntaxiques de la phrase (*Fig. IV.4.*).

Lors de l'analyse des textes, plusieurs situations se présentent où le repérage des syntagmes nominaux n'est pas toujours évident. Cela arrive parce qu'il y a des éléments anaphoriques, des ellipses, des syntagmes nominaux cachés, des syntagmes nominaux avec le déterminant zéro, etc.
Ainsi, il a fallu adopter quelques règles afin d'extraire les syntagmes nominaux de façon homogène pour obtenir des résultats statistiques cohérents dans un objectif précis : établir une grammaire de réécriture basée sur le corpus.

Une manière de résoudre ces problèmes était de s'occuper seulement de l'extraction des syntagmes de surfaces « complets » sans traitement des cas anaphoriques, élliptiques, ou cachés. Seuls les SN avec déterminant zéro sont pris en compte, car nous supposons la facilité de remédier à ce type de problème lors de l'implémentation de l'analyseur morpho-syntaxique.

Les structures syntaxiques qui ont subi cette étude sont les syntagmes nominaux complexes (les SN maximaux, en abrégé SN_max), les syntagmes nominaux simples ou inclus dans les SN_max (les SN inclus, en abrégé SN_inc), les syntagmes prépositionnels (SP), les expansions prépositionnelles (EP) et les phrases ou expressions relatives (REL).

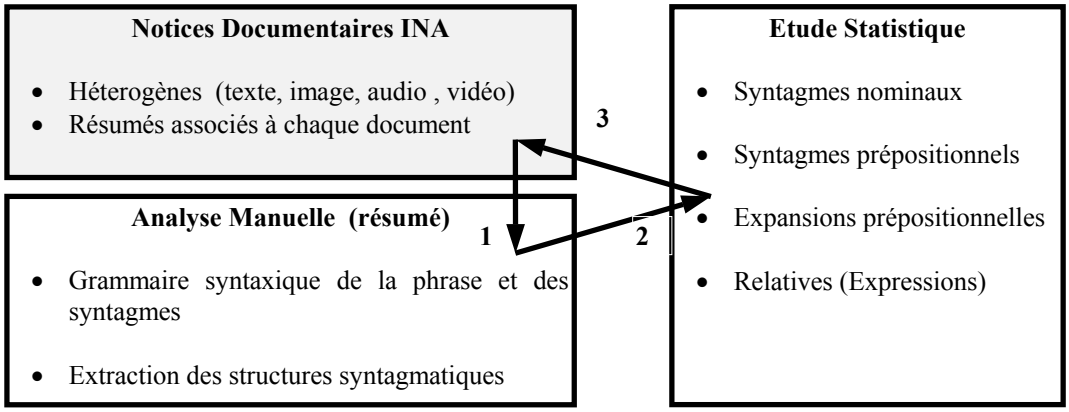


Fig. IV.4. Processus d'extraction des connaissances sur le corpus.

IV.4.1- Analyse manuelle du corpus

A- 1er Corpus : source INAthèque

TYPE de notices	Nombre de notices	Nombre de Chapeau	Nombre de Résumé	Nombre de Résumé Producteur
Notices documents audiovisuelles (programmes T.V.) :				
TOTAL	49	45	46	8
Notices documents écrits (Documents ECRITS)				
TOTAL	12	12	0	0
Notices documents audio (Radio INFO. SCIENCES) :				
TOTAL	10	10	8	0
Notices documents audio (Radio Journal Parlé SCIENCES) :				
TOTAL	13	13	11	0
TOTAL	**84**	**80**	**65**	**8**

B- 2ème Corpus : source INA-actualités

TYPE de notices	Nombre de notices	Nombre de Chapeau	Nombre de Résumé	Nombre de Résumé Producteur
Notices documents images animées (Documents SCIENCES) :				
TOTAL	35	35	35	0
TOTAL	**35**	**35**	**35**	**0**

C- 3ème Corpus : source Internet (Web)

TYPE de notices	Nombre de notices	Nombre de Chapeau	Nombre de Résumé	Nombre de Résumé Producteur
Notices documents images fixes (Animaux préhistoriques) :				
TOTAL	21	0	0	21
TOTAL	**21**	**0**	**0**	**21**

D- Corpus d'Etude : enrichissement du Corpus

TYPE de notices	Nombre de notices	Nombre de Chapeau	Nombre de Résumé	Nombre de Résumé Producteur
INAthèque				
TOTAL	**84**	**80**	**65**	**8**
INA-actualités				
TOTAL	**84**	**84**	**84**	**0**
Web				
TOTAL	**84**	**0**	**0**	**84**

IV.4.2- Résultats de l'analyse : Modèle de rédaction

Le déroulement de cette étude s'est effectué sur les champs résumés du corpus INA (résumé court (ou chapeau) et résumé détaillé (ou séquences)). Notre choix qui s'est porté sur l'étude seulement de ces champs réside au fait qu'ils sont élaborés selon des concepts méthodologiques sur l'analyse de contenu (grilles d'analyse de l'audiovisuel INA). Et par suite de cette régularité méthodologique, nous étions motivé de l'espoir de découvrir une régularité structurelle dans les textes libres des résumés.

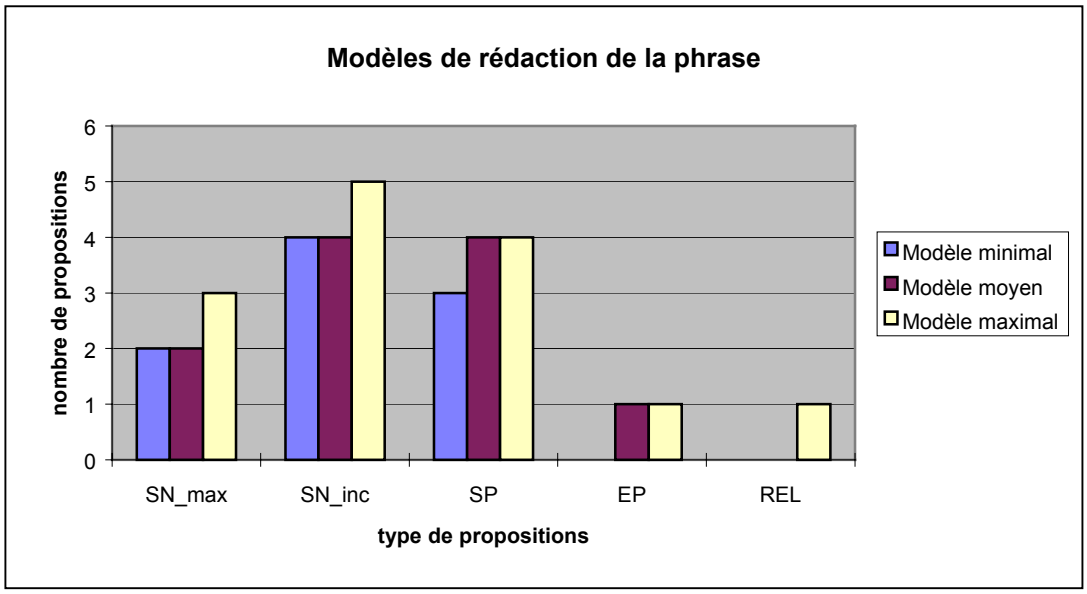
Le corpus au départ est constitué d'environ 100 notices bibliographiques de documents audiovisuels INA (notices radio et télévision) et qui a été enrichi graduellement jusqu'à environ 200 notices. Chaque notice de l'INA contient au moins deux champs résumés (enregistrements chapeau et résumé). Dans certaines notices, nous retrouvons également le résumé producteur et/ ou la description des séquences du document (enregistrements séquence par séquence).

Champs étudiés	SN_max	SN_inc	SP	EP	REL
chapeau	2.56	4.30	4.01	0.38	0.37
résumé	2.05	4.37	3.35	0.66	0.46
moyenne	**2.30**	**4.33**	**3.68**	**0.52**	**0.41**

A la suite de cette étude purement statistique, trois modèles de la phrase peuvent être envisagés :

MODELE	SN_max	SN_inc	SP	EP	REL
Modèle minimal	2	4	3	0	0
Modèle maximal	3	5	4	1	1
Modèle intermédiaire	**2**	**4**	**4**	**1**	**0**

L'analyse statistique sur le corpus de notices a révélé une stabilité grammaticale dans les descriptifs textuels (résumés). Cette révélation grammaticale cache en réalité une stabilité de rédaction des textes résumés. Nous avons observé [SIDHOM, 99a], que les documentalistes de l'INA lors de la rédaction des résumés de contenus textuels n'ont pas de contraintes rédactionnelles, structurelles ou syntaxiques pour rédiger ces résumés, si ce n'est d'appliquer la grille d'analyse de contenu des documents audiovisuels.



La réutilisation de cette stabilité grammaticale dans le sens de la production de textes permet de construire un modèle de rédaction de textes « résumés » avec une grammaire syntagmatique restreinte et bien formée. La structure syntaxique de cette grammaire est moins complexe qu'on ne l'a imaginée avant cette étude.

Nous considérons que ce modèle de grammaire syntagmatique comme modèle « cognitif » pourra nous servir à la fois comme outil d'indexation et comme celui de l'aide à la rédaction de textes [SIDHOM, 2000].

IV.4.3- Modèle de la phrase : structures internes

A- Structures préfixées PI à la phrase : *Proposition Introductive*

PI	Exemples
1. SP, ⊂ PHR	Pour les 20 ans d'AIRBUS INDUSTRIE, …
2. EP, ⊂ PHR	En parallèle, …
3. EP + SP, ⊂ PHR	En direct depuis l'observatoire de Meudon, … En compagnie de Marianne GRUNBERG-MANAGO, …
4. PPas + SP, ⊂ PHR	Embarqués à bord de l'astrolabe depuis l'extrême sud de l'Australie, …
5. PPré + SN, ⊂ PHR	Proposant un voyage à travers les sites industriels de France, …
6. Prép(en) + SNdat, ⊂ PHR	En juin 1986,
7. Prép(en) + PPré + SP, ⊂ PHR	En passant [par (la littérature)], …
8. Conj, ⊂ PHR	Cependant, …
9. Conj + Adv + SP, ⊂ PHR	Car contrairement aux Américains, …
10. « en » + PPrés, ⊂ PHR	En vaccinant,

B- Structures du SN dans la phrase : *Syntagme Nominal*

SN	Exemples
11. SN (selon la grammaire SYDO)	Ce documentaire …
12. EP ⊂ SN	Une équipe ⊃ de tournage … Un avion Hercule ⊃ de transport stratégique
13. SP ⊂ SN	La présence ⊃ d'un lac …
14. { SN, SP, EP} ⊂ SN	L'utilisation ⊃ d'images de synthèse …
15. REL (relative explicative) ⊂ SN	La présence d'un lac ⊃ qui se serait formé suite à la chute d'une météorite …
Exceptions :	
16. SN^{∇} = SN sans déterminant	Psychologues et physiciens (se penchent sur leurs multiples facettes.)

C- Structures du REL dans la phrase : *Phrase Relative*

REL	Exemples
17. /REL = Prel + SN/ ⊂ SN	… ,qui + son père, …
18. /REL = Prel + SV/ ⊂ SN	… qui + se serait formé suite à la chute d'une météorite …
19. /REL = Prel + S/ ⊂ SN	a) … qu' + il a réalisé sur le même sujet en 1973. b) … dont + le pouvoir suggestif déborde largement le cadre du bâtiment lui-même.

D- Structures du SV dans la phrase : *Syntagme Verbal*

SV	Exemples
20. V + (Prép + V-inf)+ SP	…**est** (de récupérer) de la matière cosmique
21. V + (Prép + V-inf)+ SN	…**sont montrées** (pour comprendre) les difficultés techniques et économiques
22. V + (V-inf) + SN	…**a pu** (rencontrer) AIRBUS INDUSTRIE
23. V + (V-inf) + (Prép + V-inf)+ SN	…**devait** (permettre) (d'identifier) le sexe
24. V + (PPrés) + SN	…**a suivi** (durant) trois semaines les activités d'une équipe
25. V + SN	…**sont** le reflet de notre société
26. V + SP	… **est réservée** aux avions Hercule
27. V + {SN, SP, EP, PV}	… **essaie** d'expliquer le mystère de l'étoile de Bethléem
28. V + (Adv) + SN	…**explique** (comment) les pays européens exportent des armes
29. V + (Adv) + V	…**sont** intimement **liées** …**est** ainsi **développé**
30. V + (Adv) + (Prép + V-inf)+ SN	…**s'attache** (plus) (à expliquer) la course du côté soviétique
31. V + /EP/ + SN	…**démontre** /en particulier/ la polititique de la France à ce sujet
32. V + /Conj/ + SN	…**poursuit** /donc/ cette balade à la fois historique, sociologique et architecturale
33. V	Ce chien **mord**
34. V + (Adj) + SP	…**furent** (découvertes) en 1988
35. V + {Adv, Adj}	(Il) **résout** scientifiquement

La construction de la phrase (PHR) selon notre étude s'articule autour de trois structures fondamentales, à savoir : – une structure qui précède la phrase (proposition introductive à PHR), – le syntagme nominal sujet de PHR (de nature SN complexe ou SN_max), – le syntagme verbal de S, et – la phrase relative (REL). Chacune de ces structures est identifiée en ses éléments et sous-structures composites :

$$PHR \rightarrow [PI] + SN + [REL_{SN}] + SV + [REL_{SV}]$$

$[x]$: *élément facultatif*

Nous avons noté également, la présence d'expressions type qui sont isolées à la phrase. Il s'agit de syntagmes nominaux avec le déterminant zéro, notés SN^{∇}.

IV.5- Conclusion

L'objectif recherché par les professionnels de l'information, au moment de décrire le contenu d'un document pour une constitution d'un fonds documentaire ou d'une banque de données, est d'informer les usagers. Les demandeurs de l'information retrouvent ainsi de manière pertinente les informations utiles à leur recherche. En particulier les professionnels de l'INA de France ont comme mission la constitution d'un savoir audiovisuel pour les futures générations.

Pour atteindre cet objectif, les professionnels de l'information pratiquent sur les documents un certain nombre d'opérations intellectuelles. Pour ceux de l'INA, il s'agit de l'analyse documentaire de l'audiovisuel selon la grille « d'analyse chronologique », et la constitution d'outils terminologiques et thésauriales.
L'analyse documentaire comprend deux opérations distinctes et complémentaires pour la constitution de banques de données bibliographiques (notices) :

- L'élaboration de résumés en vue de présenter à l'utilisateur une version abrégée et concise du contenu du document et de lui permettre de juger de l'opportunité de lecture, de visualisation ou de l'audition de la version source du document. Il s'agit de générer un texte court selon des critères méthodologiques bien définis.
- L'analyse réside également dans l'indexation, c'est-à-dire dans le relevé des concepts significatifs, porteurs de l'information et caractérisant le document analysé. Ce critère fait la qualité de la mémoire documentaire.

Dans les deux situations de l'activité intellectuelle (du professionnel de l'information) et de l'indexation les comportements ne sont pas les mêmes. Dans le cas de l'activité d'indexation, le professionnel est actif sur le plan linguistique et passif aux informations exprimées. Dans l'autre cas (l'activité intellectuelle), il est actif dans les connaissances exprimées, et aussi actif à générer les résumés qui incorporent une activité linguistique.

Génération de résumés, création d'outils linguistiques, mémorisation des connaissances et recherche d'informations grâce à l'indexation, sont toutes des activités intellectuelles où le traitement automatisé ou semi-automatisé par ordinateur occupent une place prépondérante.

Par rapport aux mots-clés avec lesquels on indexe habituellement les documents, le syntagme nominal est d'un côté un descripteur cohérent, car il est défini formellement, et de l'autre côté un descripteur réel et objectif, car il répond à des phénomènes de la langue.

L'intérêt de ce travail basé sur le corpus est que, sans contraintes syntaxiques ou rédactionnelles dans la production des résumés de l'INA, nous avons identifié un modèle syntaxique stable avec une grammaire simple [SIDHOM, 98]. Le modèle identifié reflète une activité intellectuelle et un comportement de rédaction homogène chez les documentalistes de l'INA.
Ce modèle, appelé « cognitif », représenté une fois, nous servira à indexer par extraction des syntagmes nominaux, et à faire de la recherche d'information basée sur les SN. L'outil pourra servir comme aide à la rédaction de textes, et pourra produire de manière automatique leur indexation.

CHAPITRE V :

Noyau d'Indexation : implémentation de l'analyseur morpho-syntaxique basée sur les automates à transitions augmentées

V.1- Introduction

Le traitement automatique du langage naturel (TALN) a progressé de manière spectaculaire avec l'apparition d'ordinateurs toujours plus puissants, et le développement parallèle de méthodes aux confluents de la linguistique, des logiques-mathématiques et de l'informatique. La recherche effectuée dans le domaine du TALN, limitée au traitement du langage écrit du français, vise tout autant à évoquer les thèmes de recherche actuels qu'à indiquer le Groupe de Recherche SYDO et le laboratoire concerné (SII enssib - Lyon).

Cette présentation reste volontairement généraliste au début puis approfondie, et cherche à fournir au lecteur une vision d'ensemble sur la mise au point d'analyseur morpho-syntaxique et des phases de traitement automatique du langage écrit.

Un préalable à la manipulation de textes écrits en français est souvent la nécessité d'effectuer une première analyse de la suite de caractères formant ces textes. Cette analyse doit accomplir les tâches suivantes:

- le regroupement des caractères formant les entités linguistiques de base [BELLOT, 97], à savoir les mots (phase de *segmentation*),
- la reconnaissance de la morphologie en ces entités, souvent grâce à un dictionnaire des formes [BOUCHART, 95] en précisant les informations morphologiques associées à chaque forme (phase d'*analyse morphologique*),
- la reconnaissance de la structure grammaticale des phrases ou de syntagmes, généralement à travers des formalismes [BARRAUD, 96] utilisant différents types de grammaires (phase d'*analyse syntaxique*).

Nous présentons, dans cette partie, les outils de modélisation du noyau d'indexation automatique. L'architecture de l'application est basée sur l'emploi et la réalisation des automates à transitions augmentées en cascade (ATN) de W. Woods [WOODS, 70-80], [BATES, 78]. Les structures objets générées par les automates sont similaires à des arbres décorés d'analyse morpho-syntaxique pour les structures de la phrase, de la phrase relative, du syntagme nominal, du syntagme prépositionnel, de l'expansion prépositionnelle, du syntagme adjectival, et du syntagme verbal.

V.2- Architecture du noyau d'indexation automatique

V.2.1- Grammaire Transformationnelle vs. Formalisme ATN

N. Chomsky en 1965 a proposé une organisation des connaissances syntaxiques en trois phases :

- **Grammaire Formelle**, permettant d'engendrer un ensemble de structures abstraites dites « structures profondes ».
  Ces structures correspondent à des phrases ou syntagmes noyaux dont on associe l'arbre de dérivation adéquate.
- **Règles de Transformations**, agissant sur les arbres de dérivation et permettant de réordonner les chaînes terminales en y ajoutant ou en supprimant des éléments. Ce qui engendre les formes possibles d'analyse de la phrase ou du syntagme.
- **Règles Morpho-phonémiques**, permettant de construire, à partir des chaînes terminales, la suite de phonèmes ou de mots correspondant à la phrase ou syntagme d'analyse.

La grammaire de la phrase ou syntagmatique complétée par le lexique forme la base du module d'analyse syntaxique qui a pour but de produire la structure profonde de la phrase (*resp.* syntagme). La structure de surface de la phrase (*resp.* syntagme) correspond au module syntaxique complété par les règles de transformations appliquées à la structure profonde de la phrase (*resp.* syntagme).

L'étude des possibilités offertes par les formalismes d'analyse de phrases ou de syntagmes a porté sur l'adoption du formalisme des Automates à Transitions Augmentées en cascade pour l'analyse du français écrit.

V.2.2- Formalisme des automates ATN (Augmented Transition Network)

Le formalisme ATN est formé de diverses classes de réseaux [WOODS, 70-86] qui correspondent aux classes des grammaires de la hiérarchie de N. Chomsky :

- Le réseau élémentaire est celui le réseau B.T.N. (Basic Transition Network) qui corresponde à l'Automate fini. Le formalisme BTN permet de définir les langages réguliers de manière plus concise que les automates finis. Cette concision est obtenue en faisant opérer aux catégories syntaxiques d'une grammaire des rôles de symboles d'entrée ou de terminaux pour le réseau, alors que dans un automate fini les seuls symboles d'entrée sont des mots du vocabulaire.
- Le réseau RTN (Recursive Transition Network), qui occupe la seconde place de la hiérarchie des réseaux, est obtenu à partir d'un réseau BTN et de l'adjonction de la récursivité dans le réseau. Le RTN correspond à l'automate à pile et accepte les langages hors-contextes.
- Le réseau ATN (Augmented Transition Network) figure au sommet de la hiérarchie des réseaux. Il est obtenu à partir du réseau RTN et d'un certain nombre d'ajouts (ou augmentations) permettant d'y intégrer l'équivalent de traitements réalisés par les grammaires transformationnelles. L'ATN est équivalent dans sa puissance de traitements symboliques à la Machine de Turing. A chaque réseau ATN est attaché un ensemble de registres précisant les *attributs* (exemple : genre et nombre pour un groupe syntaxique, etc.) et les *rôles* attachés à chaque structure engendrée (exemple, position sujet ou complément d'objet pour le groupe syntaxique).

Dans les réseaux ATN, des conditions et des actions sont associées aux différents arcs du réseau et qui vont permettre toutes ces extensions. Les conditions servent à restreindre les circonstances pour emprunter un arc dans un état du réseau : les conditions sur les propriétés du mot ou du constituant correspondant à l'arc, mais aussi des constituants déjà trouvés. Les actions serviront à construire des descriptions partielles sur les parties analysées. Ces descriptions seront contenues dans des registres attachés à chaque réseau.

V.3- Implémentation de l'analyseur morpho-syntaxique

V.3.1- Spécification de la syntaxe du formalisme ATN

Une spécification (en notation B.N.F.) utilisée par W.A. Woods (1980) permet de décrire la syntaxe du formalisme des ATN. C'est une version améliorée par rapport à celle spécifiée par l'auteur en 1970. Les changements majeurs dans cette dernière spécification sont d'ordre formel permettant une grande flexibilité de factorisation et de partage commun entre les différentes parties de phrases types à analyser [WOODS, 80].

<ATN> : := (<machinename> (accepts <phrasetype>*) <statespec>*)
; un ATN est une liste de nom de machines (automates), une spécification de phrases types qui seront acceptées par l'ATN, et une liste de spécification des états.

<statespec> : := (<statename> { optional <initialspec> } <arc>*)
; la spécification d'un état dans l'automate consiste à l'attribution d'un nom à cet état et des arcs de transition, et en option s'il s'agit d'un état initial.

<initialspec> : := (initial <phrasetype>*)
; indique un état initial pour une phrase type indiquée.

<arc> : := (<phrasetype> <nextstate> <act>*)
; une transition qui consomme une phrase type indiquée.
: := (<pattern> <nextstate> <act>*)
; une transition qui consomme un élément en entrée et qui vérifie un modèle <pattern>.
: := (J <nextstate> <act>*)
; une transition vers un nouvel état sans consommer un élément d'entrée.
: := (POP <phrasetype> <form>)
; indique un état final pour une phrase type indiquée et spécifie une forme à retourner à sa structure.

<nextstate> : := <statename>
; spécifie l'état suivant pour une transition.

<pattern> : := (<pattern>*)
; vérifie une liste dont les éléments correspondent à la succession de pattern spécifié.
: := <wordlist>
; correspond à tout élément dans la liste.
: := ε
; correspond à n'importe quel élément.
: := --
; correspond à n'importe quelle sous-séquence.
: := <form>
; correspond à une valeur de <form>.
: := <<classname>>
; correspond à n'importe quel élément qui a ou hérite les traits de la classe.

<wordlist> : := { ' <word> | ' <word>, <wordlist> }

<act> : := (transmit <form>)
; transmettre la valeur de <form> comme une sortie.
: := (setr <registername> <form>)
; fixe le registre à la valeur de <form>.

: := (addr <registername> <form>)
; ajouter la valeur de <form> à la fin de liste dans le registre indiqué
; (initialement a pour valeur NIL quand le registre n'est pas fixé).
: := (require <proposition>)
; abandonner le chemin si la proposition est fausse.
: := (dec <flaglist>)
; fixe la liste des booléens indiqués.
: := (req <flagproposition>)
; abandonner le chemin si la proposition est fausse.
: := (once <flag>)
; équivalent à (req (not <flag>)) (dec <flag>).

<flagproposition> : := <boolean combination of flag registers>

<propostion> : := <form>
; la proposition est fausse si la valeur de <form> est NIL.

<form> : := !<registername>
; retourne les contenus du registre.
: := ′ <liststructure>
; retourne une copie de la structure en liste à l'exception de toutes les expressions précédées par ! sont remplacées par leur valeur et celles précédées par @ ont leur valeur insérée comme une sous-liste.
: := !c
; contenus des constituants courants du registre.
: := !<liststructure>
; retourne la valeur de la structure en liste interprétée comme une expression fonctionnelle.

V.3.2- Implémentation de l'analyseur en Langage objet

L'analyse syntaxique est une étape importante dans l'analyse linguistique de textes. Elle permet d'organiser les mots d'une phrase en groupe de mots, explicitement en syntagmes spécifiques, et de mettre les relations entre les groupes constituant cette phrase. Le résultat de l'analyse dans notre travail sera l'analyse du corpus, l'indexation de contenus textuels, le traitement de requêtes (en langage naturel) en vue de la recherche d'informations.

V.3.2.1- Segmentation superficielle

Avant toute application d'un formalisme linguistique, le premier traitement qui s'opère sur un texte *vs.* une phrase d'entrée consiste en la segmentation en séquences de mots ou *formes*.
Les opérations invoquées par ce type de traitement y compris la régularisation de surface portent sur les mots superficiels qui composent les phrases du texte. Ces opérations précèdent même l'analyse morpho-syntaxique, car les traitements d'analyse morphologique et lexicale ont pour objets des mots.

- Découpage de texte en phrases

Les frontières de phrases dans un texte sont repérables en dehors de toute analyse morpho-syntaxique. Le parcours de la chaîne de caractères représentant le texte et le repérage des marques de fin de phrase permettent d'y parvenir au découpage. Le résultat de ce découpage donne une suite de chaînes de caractères dont chacune représente une phrase.
Les ponctuations fortes (marqueurs de fin de la phrase) servant à cette première transformation sont : /./, /;/, /?/, /!/. Nous avons observé lors de l'étude du corpus que certaines ponctuations comme : /:/ et /-/ ont servi de ponctuation forte (marqueur de début de la phrase).

- Découpage en formes

L'objet de cette étape de transformation est de faciliter les tâches d'analyse ultérieure. La lecture de la phrase revient à la lecture d'une liste de formes graphiques. Un certain nombre de règles sont appliquées :

1- l'occurrence d'un ou plusieurs espace(s) pour séparer les occurrences de formes.
2- l'occurrence de certaines marques de ponctuations doubles : /«/, /»/, /(/, /)/, /[/, /]/, /‘/, /’/, /{/, /}/, accolées au mot en amant et en aval, pour séparer l'occurrence du mot.
3- l'occurrence des formes élidées (ou présence de l'apostrophe) : des règles de transformation sont appliquées pour restituer les formes originelles ; /d'/ en /de/, /j'/ en /je/, /m'/ en /me/, /n'/ en /ne/, /t'/ en /te/, /s' (+il ou +ils)/ en /si/ sinon en /se/, /l'on/ en /on/, /c'/ en /ce/, /l'/ en /le/ ou /la/ ou /l'/ selon le contexte.
4- L'occurrence du trait d'union est supprimé dans les contextes suivants : /xxx-[je..il]/ en /xxx/ /[je..il]/* avec la marque particule préverbale sujet pour *, /xxx-[ci, là]/ en /xxx/ /[ci, là]/, /xxx-t-yyy/ en /xxx/ /yyy/.
5- L'occurrence de locutions prépositionnelles ou de mots composés est une considération à retenir lors de l'application des règles de découpage où l'espace, l'apostrophe et le trait d'union font partie intégrante de la forme. La liste de ces formes reste à établir : /à cause de/, /grand-père/, /presqu'île/, etc.

Certaines contraintes d'application de ces règles dépendront des traitements ultérieurs ou bien seront différer à des niveaux d'analyse supérieures comme la phase d'analyse syntaxique.

V.3.2.2- Classification des mots

La segmentation superficielle décrite ci-dessus est un premier pas vers le découpage du texte. La décomposition du texte en phrases, de la phrase en ses formes et l'élimination des formes élidées est une première régularisation du texte. D'autres traitements de régularisation doivent s'effectuer et ont pour but de réduire l'ensemble des formes, de limiter le nombre de catégories d'analyse.
Le modèle classificatoire des mots que nous allons décrire et qui va servir à effectuer l'analyse morpho-synatxique a été proposé pour l'essentiel par A. Berrendonner puis repris et modifié par J.-P. Metzger. La solution proposée dans ce modèle consiste à se donner un nombre très restreint de catégories syntaxiques, chacune ayant un comportement distributionnel bien défini.

A- Liste des catégories syntaxiques

Catégories principales	Modèle de J.-P. Metzger	Modèle d'A. Berrendonner
Catégories nominales :	F : noms-adjectifs	F
	D : prédéterminants	D
	P : prépositions	P
Catégories intermédiaires :	C : coordonnants	C
	W : adverbes	W
	T : ponctuations	T
Catégories verbales :	V : verbes	V
	Y : pronoms préverbaux	Y
	Q : subordonnants	Q
Catégories « modifiées » :	W-NEG (partie de W)	G : Mots négatifs
	F-NOM-PRO (partie de F)	H : pro-phrases

B- Sous-catégorisation syntaxique

L'analyse en constituants de la phrase nécessite une classification plus fine que celle induite par les neuf catégories principales. Certaines catégories peuvent ainsi être subdivisées en sous-catégories établies sur des bases distributionnelles.

Catégorie	Sous-catégorie		Variables	Valeurs
F	NOM (noms) :	COM (communs)	GR (genre)	MAS (masculin) FEM (féminin) GRN (non marqué)
		PRO (pronoms)		
		PRP (propres)		
			NB (nombre)	SIN (singulier) PLU (pluriel) NBN (non marqué)
			PE (personne)	PE1 (je) .. PE5(ils)
			DQ (dérivation)	AGE (déverbaux agents) PPA (participes passés) PPR (participes présents) DVB (dérivés de verbes) DAJ (substantifs dérivés d'adj.)
	ADJ (adjectifs)		CI (continu/discret)	CTN (continu) DCT (discontinu) CIN (non marqué)
	NAN (adjectifs et noms)		AN (animé/inanimé)	ANI (animé) INA (inanimé) ANN (non marqué)
D	DEF (définis)			
	NUM (numéraux cardinaux)			
	IND (indéfinis et autres)			
W	AAJ (modificateur d'adjectif)			
	QUA (de quantité)			
	PRO (déictiques et anaphoriques)			
	TAM (temps, aspect et mode)			
P			RV (rection verbale)	SCO (sans complément)
				ACC (complément accusatif)
				DAT (complément datif)
				LOC (complément locatif)
				ABL (complément ablatif)
				ADA (accusatif + datif)
				ALO (accusatif + locatif)
				AAB (accusatif + ablatif)
				DAB (datif + ablatif)
C				
V			NB PE RV	
Y	IN1 (pré-verbaux Sujets) : {je, tu, il, elle, ils, elles, on, ce}		GR NB PE	
	IN2 (pré-verbaux Compléments) : {le, la, l', les, lui, leur, me, te, se, en, y}			
	INN (pré-verbaux Neutres) : {nous, vous}			
Q				
T			VP (valeur de ponctua.)	VP1 (ponctuation faible) VP2 (ponctuation forte) VPN (ponctuation non marqué)

C- Modèle linguistique

Le recours à la linguistique est une hypothèse de travail sans quoi l'analyse d'un énoncé en langue naturelle ne peut s'opérer sans faire appel à des fondements linguistiques. Ce recours est le seul guide logiquement opératoire dans le passage des formes de surfaces au codage recherché : condition nécessaire et suffisante pour catégoriser, regrouper et interpréter les énoncés.

Le modèle linguistique adopté, celui du groupe SYDO, a une structure modulaire. Les régularisations de surface s'opèrent avant l'analyse morphologique. Cette structure repose sur le fait qu'il est plus facile de concevoir et de maintenir des grammaires dont le rôle est précisément défini.

Dans ce cadre, la classification relève entièrement du modèle linguistique et est indépendante du mode de calcul choisi. La fonction classificatoire est de nature lexicale consistant à attribuer à chaque forme du texte analysé une *catégorie syntaxique* et une *fonction flexionnelle* [BERRENDONNER, 90]. Le tableau (*Tab.V.3.2.2.*) qui suit permet de positionner la forme dans la classification :

Variables syntaxiques		**Variables flexionnelles**		**Variables lexicales**	
Variable	**Valeur**	**Variable**	**Valeur**	**Variable**	**Valeur**
NA (Nominaux)	NOM ADJ NAN	**GR** (Genre)	MAS FEM GRN	**NN** (s/type Nominal)	PRP COM PRO
PA (Participes)	PPA PPR	**NB** (Nombre)	SIN PLU NBN	**CI** (Discrétion)	CTN DCT CIN
IN (Préverbaux)	IN1 IN2 INN	**PE** (Personne)	PE1(je) PE2(tu) PE3(il, ils, on) PE4(nous) PE5(vous)	**AN** (Animation)	ANI INA ANN
PS (Préposition sous-jacente IN)	PDE PA1 PA2			**DQ** (non-Dérivé)	DVB DAJ AGE
VB (Type Verbal)	INF FIN			**VA** (valence verbale)	VA0 VA1 VA2 VA3 VA4
VX (Auxiliarité)	AUX ORD			**AP** (Opérativité)	APO AAU
				NU (Quantification)	NUM NNU
				NG (Polarité Négative)	NEG NNG
				VW (Types d'Adverbes)	AVJ TAM NUM PRO
				VP (Valeur de Ponctuation)	VP1 VP2 VP3
				FF (Type Séquentiel)	FOR FAI FFN

Tab.V.3.2.2. Modèle Classificatoire : inventaire des trois classes (Berrendonner, 1983).

V.3.2.3- Organisation et représentation des données lexicales

L'analyse d'une forme dans le texte consiste à la positionner dans la classification qui vient d'être présentée. Il s'agit de lui affecter des traits grammaticaux correspondant à une catégorie et des valeurs de sous-catégorisation d'après le modèle de définitions syntaxique, flexionnelle et lexicale. Pour un mot ou une forme composite donné(e), ses traits ne peuvent être déterminés que par la consultation d'un lexique. Un lexique que l'on peut définir et construire comme un ensemble organisé de mots et de traits grammaticaux (catégories, sous-catégories, et valeurs de variables) selon le modèle classificatoire.
Dans cette perspective, une base de donnée a été construite, en introduisant les formes possibles de la langue avec toutes leurs analyses possibles. Cette base nous a servi comme un dictionnaire électronique, où l'analyse d'une forme revient à une consultation automatique du dictionnaire. Une telle solution réside dans sa simplicité de mise en oeuvre, mais a nécessité la création d'un dictionnaire de taille importante. En général, un nom a deux formes distinctes, un adjectif quatre formes et un verbe trente formes environ. Le dictionnaire adapté à cette solution a atteint 400.000 entrées.

V.3.2.4- Régularisation morpho-syntaxique

Le modèle classificatoire ci-dessus ne prend pas en compte l'ensemble des comportements possibles. Mais conformément à la stratégie qui gouverne le modèle linguistique, la préférence était de ramener un cas particulier à un cas général. L'avantage de cette stratégie est de ne pas augmenter artificiellement le nombre de modèles.
Ainsi, pour le fonctionnement du modèle, chaque liste de formes représentant une phrase segmentée du texte est soumise à une pré-analyse. Cette pré-analyse a pour effet de décomposer certaines formes de la liste en des séquences de formes catégorisables et qui s'intègrent toutes dans la classification. Les règles se subdivisent en plusieurs groupes chacun opérant après l'autre :

1. **Eclatement d'amalgames orthographiques :**
/lequel/ en /le,D/+/quel/, /lesquels/ en /les,D-MAS/+/quels/, /auquel/ en /au/+/quel/, etc.
2. **Eclatement d'amalgames orthographiques avec inversion :**
/là-dedans/ en /dans,P-FAI/+/là,W-FOR/, /ci-dessous/ en /sous,P-FAI/+/ici,W-FOR/, etc.
3. **Eclatement d'amalgames morphologiques :**
/au/ en /à/+/le,D/, /aux/ en /à/+/les,D/, /du/ en /de,P/+/le,D/, /des/ en /de,P/+/les,D/, etc.

Les détails de la liste complète des transformations sont à consulter dans les travaux de : [METZGER, 88] et [LALLICH, 90].

V.3.2.5- Analyse flexionnelle

A l'issue du traitement effectué par la régularisation morpho-syntaxique, la phrase en cours d'analyse se présente en une liste de formes dont bon nombre sont déjà analysées. Certaines l'ont été au cours de la décomposition de l'amalgame d'autres au cours de l'analyse lexicale partielle. Afin que l'analyse soit complète et engager l'analyse syntaxique proprement dite, il est nécessaire de compléter l'analyse des formes restantes.

Etant donné que la base de données lexicales renferme les formes fléchies, chacune d'elle devrait normalement comporter son profil syntaxique (ensemble de ses traits syntaxiques), son profil lexical (ensemble de ses traits lexicaux) et tout ou partie de son profil flexionnel

(ensemble de ses traits flexionnels). Le cas échéant, certains des traits flexionnels doivent être déterminés directement à partir de la forme elle-même.
L'analyse d'une occurrence de mot du texte consiste alors à retrouver dans le lexique l'ensemble des traits associés au mot correspondant, sinon (mot inexistant dans le lexique) à calculer sa base ou sa forme réduite afin d'en déduire son profil syntaxique, lexical et flexionnel.

A- Analyse d'une occurrence de mot

Lors de l'analyse d'une occurrence de mot (ou forme), trois cas peuvent se présenter :

1. Le mot est déjà analysé : reconnu comme tel ;
2. Le mot n'est pas analysé et il se trouve tel quel dans le lexique : extraire son profil syntaxique, lexical et flexionnel ;
3. Le mot n'est pas analysé et ne figure pas dans le lexique : considéré comme une forme fléchie et doit être soumise à une analyse flexionnelle ;
4. Le mot a subi une analyse flexionnelle et sans succès : considéré comme une forme de nom-propre (F-NOM-PRP).

B- Algorithme d'analyse flexionnelle

Les mots appartenant aux catégories F, V, Y, D sont porteurs des marques de genre, nombre et personne, aussi bien le temps et le mode pour les formes verbales. Or, le lexique construit dans la base de données dispose, pour 10.000 verbes avec les formes conjuguées et de ~100.000 formes non-verbales avec des informations s'y rapportant aux marques flexionnelles. Par la richesse de notre lexique, nous nous limiterons donc, à l'analyse des formes nominales ou adjectivales (F-NOM/ F-ADJ/ F-PRP).

L'algorithme consiste à déterminer à partir de la forme ses traits flexionnels associés à la base (genre et nombre), puis d'effectuer la recherche dans le lexique (_Recherche_Dico) avec la base et d'en compléter son profil selon les traits associés (_Analyse).

Algorithme itératif de la fonction Recherche_Dico(forme) pour l'analyse flexionnelle :

0. si (Recherche_Dico(forme) avec succès)
        Analyse(forme) ;
   sinon
        forme = Base + flexion_genre + flexion_nombre ;

1. si (Recherche_Dico(forme – flexion_nombre) avec succès)
        Analyse(forme – flexion_nombre) ;
   sinon

2. si (Recherche_Dico(forme – flexion_genre) avec succès)
        Analyse(forme – flexion_genre) ;
   sinon

3. si (Recherche_Dico(forme – flexion_genre – flexion_nombre) avec succès)
        Analyse(forme – flexion_genre – flexion_nombre) ;
   sinon

4. si ((forme = nom_prpore) avec succès)
        Analyse(forme) ;

C- Levée des ambiguïtés

A l'issue des traitements de la séquence superficielle et avant la mise en oeuvre des procédures de reconnaissance de certaines classes de séquences (syntagmes, propositions, et phrases), il est nécessaire de procéder à d'autres opérations morpho-syntaxiques : les analyses parasites attribuées à certaines occurrences de mots ambigus.

Les mots considérés hors contexte peuvent contenir plusieurs analyses. La levée d'une ambiguïté s'appuie sur le contexte des mots qui précèdent et ceux qui suivent l'occurrence concernée.
Une méthode purement linguistique s'appuie sur un système grammatical. S'en suivent l'obligation et l'interdiction (*Fig.V.3.2.5.*) de certaines occurrences dans des séquences de catégories déterminées (ou sous-catégories). Cette solution « heuristiques » liée aux problèmes d'ambiguïtés ne garantit pas une résolution totale. Il s'agit dans ces principes heuristiques d'étudier l'occurrence plausible du mot courant en fonction du mot précédant et du mot suivant. Cette opération fait appel à une activation de certaines catégories (ou sous-catégories) et la désactivation du reste.

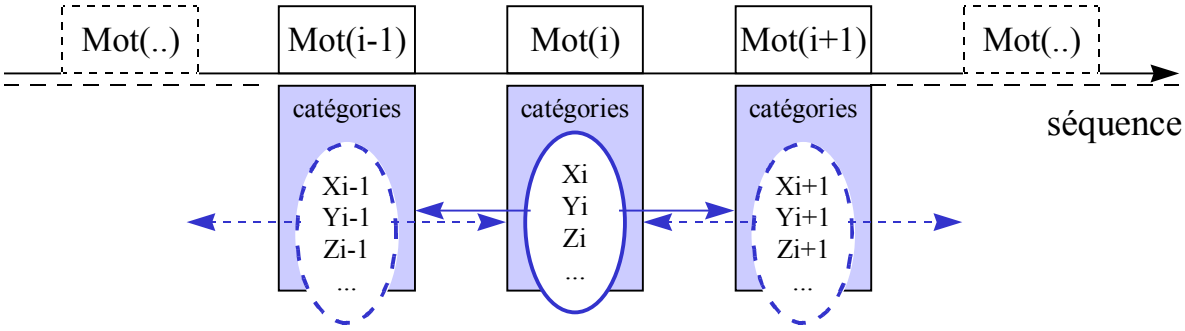


Fig.V.3.2.5. Heuristiques d'élimination des ambiguïtés.

A partir de ce modèle heuristique, nous arrivons à réduire considérablement l'ambiguïté sur la phrase et réduire le temps de traitements lors de son analyse syntaxique. Nous donnons la liste de ces règles heuristiques.
$\{+X\}_{\text{Précéd/cour/suiv}}$ représente la catégorie syntaxique X dans la liste des catégories du mot_précédant/ mot_courant/ mot_suivant. {–X} représente le masquage de cette catégorie.

➢ Modèle heuristique sur le mot_précédant

$$1: \quad \{+D\}_{\text{Précéd}} \wedge \{+P\}_{cour} \rightarrow \{+D\}_{précéd} \wedge \{-P\}_{cour}$$
$$2: \quad \{+D\}_{\text{Précéd}} \wedge \{+V\}_{cour} \rightarrow \{+D\}_{précéd} \wedge \{-V\}_{cour}$$
$$3: \quad \{+D\}_{\text{Précéd}} \wedge \{+C\}_{cour} \rightarrow \{+D\}_{précéd} \wedge \{-C\}_{cour}$$
$$4: \quad \{+A\}_{\text{Précéd}} \wedge \{+D\}_{cour} \rightarrow \{+A\}_{précéd} \wedge \{-D\}_{cour}$$
$$5: \quad \{+N\}_{\text{Précéd}} \wedge \{+D\}_{cour} \rightarrow \{+N\}_{précéd} \wedge \{-D\}_{cour}$$
$$6: \quad \{+N\}_{\text{Précéd}} \wedge \{+W\}_{cour} \rightarrow \{+N\}_{précéd} \wedge \{-W\}_{cour}$$
$$7: \quad \{+P\}_{\text{Précéd}} \wedge \{+C\}_{cour} \rightarrow \{+P\}_{précéd} \wedge \{-C\}_{cour}$$
$$8: \quad \{+P\}_{\text{Précéd}} \wedge \{+V\}_{cour} \rightarrow \{+P\}_{précéd} \wedge \{-V\}_{cour}$$
$$9: \quad \{+V\}_{\text{Précéd}} \wedge \{+N\}_{cour} \rightarrow \{+V\}_{précéd} \wedge \{-N\}_{cour}$$

➢ **Modèle heuristique sur le mot_suivant**

1 : $\{+P\}_{cour} \wedge \{+D\}_{suiv} \rightarrow \{-P\}_{cour} \wedge \{+D\}_{suiv}$

2 : $\{+Y\}_{cour} \wedge \{+A\}_{suiv} \rightarrow \{-Y\}_{cour} \wedge \{+A\}_{suiv}$

3 : $\{+V\,\text{inf}\}_{cour} \wedge \{+A\}_{suiv} \rightarrow \{-V\,\text{inf}\}_{cour} \wedge \{+A\}_{suiv}$

4 : $\{+Y\}_{cour} \wedge \{+N\}_{suiv} \rightarrow \{-Y\}_{cour} \wedge \{+N\}_{suiv}$

5 : $\{+V\}_{cour} \wedge \{+N\}_{suiv} \rightarrow \{-V\}_{cour} \wedge \{+N\}_{suiv}$

6 : $\{+W\}_{cour} \wedge \{+N\}_{suiv} \rightarrow \{-W\}_{cour} \wedge \{+N\}_{suiv}$

7 : $\{+Y\}_{cour} \wedge \{+P\}_{suiv} \rightarrow \{-Y\}_{cour} \wedge \{+P\}_{suiv}$

8 : $\{+W\}_{cour} \wedge \{+P\}_{suiv} \rightarrow \{-W\}_{cour} \wedge \{+P\}_{suiv}$

9: $\{+W\}_{cour} \wedge \{+V\}_{suiv} \rightarrow \{-W\}_{cour} \wedge \{+V\}_{suiv}$

10 : $\{+D\}_{cour} \wedge \{+V\}_{suiv} \rightarrow \{-D\}_{cour} \wedge \{+V\}_{suiv}$

11 : $\{+D\}_{cour} \wedge \{+Y\}_{suiv} \rightarrow \{-D\}_{cour} \wedge \{+Y\}_{suiv}$

12 : $\{+C\}_{cour} \wedge \{+Y\}_{suiv} \rightarrow \{-C\}_{cour} \wedge \{+Y\}_{suiv}$

➢ **Modèle heuristique sur le mot_courant**

1 : $\{+A,+Vppas\}_{cour} \wedge \{+F\}_{suiv} \rightarrow \{+A\}_{cour} \wedge \{+F\}_{suiv}$

2 : $\{+F\}_{cour} \wedge \{+A,+Vppas\}_{suiv} \rightarrow \{+F\}_{cour} \wedge \{+A\}_{suiv}$

3 : $\{+D\}_{cour} \wedge \{+F\}_{suiv} \rightarrow \{+D\}_{cour} \wedge \{+N\}_{suiv}$

Comparativement à notre modèle heuristique, il existe des algorithmes de levée des ambiguïtés qui évaluent sur une phrase la probabilité de chaque suite de catégories. Par exemple, un algorithme basé sur le modèle Markovien d'ordre 2 donne d'excellents résultats [KALLAS, 87]. A titre d'illustration, nous donnons un schéma du fonctionnement du modèle de Markov à l'ordre 1 (*Fig.V.3.2.5.*):

Levée des ambiguïtés par un Modèle de Markov d'ordre 1

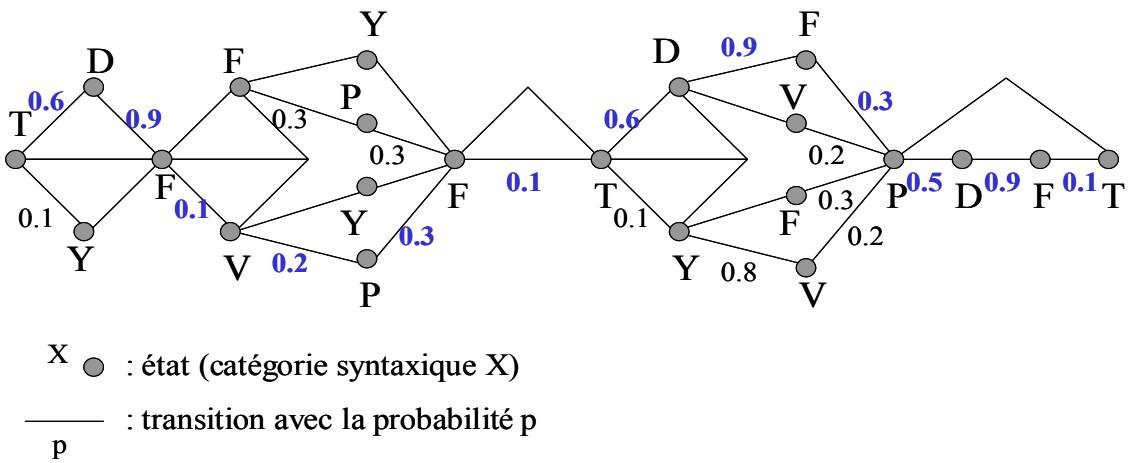


Fig.V.3.2.5. : Schéma de fonctionnement d'un Modèle de Markov ([LALLICH, 90]p.47).

V.3.2.6- Grammaire de réécriture de la phrase et des syntagmes

Cette grammaire a été proposée, pour l'essentiel, par A. Berrendonner pour l'analyses des syntagmes nominaux. Elle a servi de support à plusieurs travaux de recherche effectués dans le cadre du groupe SYDO. Nous avons également retenu les améliorations apportées sur cette grammaire par les travaux de M. Le Guern [LE GUERN, 91] [LE GUERN, 94 a et b], J.-P. Metzger [METZGER, 85] (*réécriture du syntagme nominal*), M. De Brito [DE BRITO, 91] (*reconnaissance du syntagme nominal*), O. Larouk (*traitement de la coordination*), M. Chawk (*réécriture du déterminant complexe D'*) [CHAWK, 93].

L'écriture de cette grammaire est inspirée de la notion X-barre, de N. Chomsky, pour représenter les structures syntaxiques organisées hiérarchiquement à partir de catégories principales : $X, \overline{X}, \overline{\overline{X}}$ (ou X, X', X''). Employer cette notion permet de générer les syntagmes principaux par l'emploi des deux règles :

$$(i):\ \overline{\overline{X}} \rightarrow \ spéc.\overline{X}\,\overline{X} \quad et \quad (ii):\ \overline{X} \rightarrow \ X$$

A- Les règles du syntagme nominal

Description	N° règle	Règle
syntagmes nominaux :	1	N''→ D' + N + F-PRP
	2	N''→ D' + N'
	3	N''→ NOM-PRO
	4	N''→ NOM-PRP
expressions nominales :	5	N'→ N + SP + (SP)
	6	N'→ N + A'
	7	N'→ N + CI
	8	N'→ N + LA
	9	N'→ N
expressions prédéterminatives :	10	D'→ D-DEF + D-NUM
	11	D'→ P-DE + D-DEF
	12	D'→ W-QUA + P-DE + D-DEF
	13	D'→ W-QUA + P-DE
	14	D'→ D
centres adjectivaux :	15	A'→ W-AAJ + A
	16	A'→ A + EP
	17	A'→ F-ADJ,REL
	18	A'→ A
centres nominaux :	19	N→ N + EP
	20	N→ N + A(QUA)
	21	N→ A(QUA) + N

nominaux :	22	N→ F-NOM
	23	N→ F-NAN
	24	A→ F-NAN,(QUA)
	25	A→ F-ADJ,(QUA)
syntagme prépositionnel :	26	SP→ P + N''
expansion prépositionnelle :	27	EP→ P + N'
coordination des adjectifs : - propriétés communes (QUA1)=(QUA2)	28	A(QUA)→ A(QUA1) + C + A(QUA2)
- propriétés différentes (QUA1)≠(QUA2)	29	A_C(QUA)→A(QUA1) + C + A(QUA2)

- Cas particuliers des règles SN

Description	N° règle	Règle
SN avec D'-zero : $SN^{\underline{z}}$	1	N'' → ∅ + N'
Syntagme date : $SN^{\underline{d}}$	2	N''→ D' + K + N' + (K)
Syntagme mesure : $SN^{\underline{m}}$	3	N''→ K + N'
Expression quantitative : K	4	K → unité-de-mesure

B- Les règles du syntagme verbal

Description	N° règle	Règle
Syntagme verbal :	1	SV→ Y + V"
	2	SV→ V"
	3	SV→ V" + N"
	4	SV→ V" + SP
Expressions verbales :	5	V"→ V' + F-ADJ
	6	V"→ V' + W
	7	V"→ V' + EP
	8	V" → V' + W + PPAS
	9	V" → V' + P + VINF
	10	V" → V' + W + P + VINF
	11	V" → Y* + V'
	12	V" → V'
Centres verbaux :	13	V' → V + V
	14	V' → V

C- Les règles de la Phrase relative

Description	N° règle	Règle
Phrase relative :	1	REL → F-PRO-REL + N''
	2	REL → F-PRO-REL + SV
	3	REL → F-PRO-REL + N'' + SV

D- Les règles de la proposition introductive de la phrase

Description	N° règle	Règle
Proposition introductive :	1	PI → SP
	2	PI → EP
	3	PI → EP + SP
	4	PI → A' + SP
	5	PI → A' + N''
	6	PI → P-en + $SN^{\underline{d}}$
	7	PI → P-en + PPRES + SP
	8	PI → P-en + PPRES
	9	PI → C
	10	PI → C + W + SP

E- Les règles de la phrase

Description	N° règle	Règle
phrase :	1	PHR → N'' + SV
	2	PHR → PI + N'' + SV
	3	PHR → (PI) + N'' + (REL) + SV
	3'	PHR → (PI) + N'' + SV + (REL)

F- Les règles de coordination majeures

Description	N° règle	Règle
coordination :	1	$N''_C \rightarrow N'' + C + N''$
	2	$SV_C \rightarrow SV + C + SV$
	3	$V_C \rightarrow V + C + V$
	4	$A_C \rightarrow A + C + A$
	5	$W_C \rightarrow W + C + W$
	6	$EP_C \rightarrow EP + C + EP$
	7	$SP_C \rightarrow SP + C + SP$
	8	$PHR_C \rightarrow PHR + C + PHR$

G- L'accord en genre, en nombre et en personne

Une bonne reconnaissance des occurrences de SN et de propositions au sein d'une phrase implique, non seulement le repérage de certaines séquences de catégories (ou sous-

catégories), mais aussi la prise en compte des restrictions sélectives qui pèsent sur l'association de certaines unités dans les constructions correspondantes :

1. Les contraintes d'accord en genre et en nombre s'exercent soit dans un noeud N'' : accord prédéterminant-nom, soit dans un noeud N : accord nom-adjectif. D'autre part, une séquence de catégorie N ou N'' incluse par l'intermédiaire d'un SP ou EP, dans une autre catégorie N ou N'' n'est pas soumise aux contraintes d'accord.

Description	N° règle	contraintes
syntagmes nominaux :		
N''→ D' + N'	1	genre(N'') = genre(D') = genre(N') nombre(N'') =nombre(D') = nombre(N')
D'→ D	2	genre(D') = genre(D) nombre(D') = nombre(D)
N'→ N	3	genre(N') = genre(N) nombre(N') = nombre(N)
genre(X) → GR	4	GR = valeur dans {MAS, FEM, GRN}
nombre(X) → NB	5	NB = valeur dans {SIN, PLU, NBN}
accord_GR(X,Y) → GR	6	genre(X) = genre(Y) = GR
accord_NB(X,Y) → NB	7	nombre(X) = nombre(Y) = NB

2. Les contraintes d'accord en nombre et en personne s'exercent entre un noeud N'' et un noeud V'' (dans un noeud SV). Dans le cas de pronom préverbal (sujet) Y dans un noeud SV, les mêmes contraintes s'exercent entre Y et V''. En général, ces deux principales contraintes se retrouvent dans la structure de la phrase ou de la phrase relative.

Description	N° règle	contraintes
Phrases :		
PHR → N'' + SV	1	nombre(PHR)= nombre(N'')= nombre(SV) personne(PHR)= personne(N'')=personne(SV)
REL → F-PRO-REL + N'' + SV	2	nombre(REL)= nombre(N'')= nombre(SV) personne(REL)= personne(N'')=personne(SV)
syntagmes verbaux :		
SV→ Y + V"	3	nombre(SV)= nombre(Y)= nombre(V'') personne(SV)= personne(Y)=personne(V'')
SV→ V", V"→ V' et V'→ V	4	nombre(SV)=nombre(V'')=nombre(V')=nombre(V) personne(SV)=personne(V'')=personne(V')=personne(V)
Syntagme nominal :		
N''→ D' + N'	5	nombre(N'') =nombre(D') = nombre(N') personne(N'')=personne(D')=personne(N')
nombre(X) → NB	6	NB = valeur dans {SIN, PLU, NBN}
personne(X) → PE	7	PE = valeur dans {PE1, PE2, PE3, PE4, PE5}
accord_NB(X,Y) → NB	8	nombre(X) = nombre(Y) = NB
accord_PE(X,Y) → PE	9	nombre(X) = nombre(Y) = PE

H- Rattachement des syntagmes prépositionnels

Selon J.-P. Metzger, le problème du rattachement d'un SP se pose en ces termes :

« étant donné un SP précédé de plusieurs verbes et/ou nominaux, à laquelle de ces unités doit-il être rattaché comme complément ? ».

Ce rattachement est important pour une analyse en constituants, car il conditionne l'appartenance du SP en question au syntagme (nominal, adjectival ou verbal) et qui ne peut être prédite par les seules règles de réécriture.
La règle suivante, formalisée par son auteur [METZGER, 88], et fondée sur les capacités rectionnelles des verbes et nominaux, paraît à reproduire le meilleur taux d'analyses satisfaisantes :

1. Tout SP est rattaché au nominal ou verbe le plus proche parmi ceux qui sont susceptibles de le régir ;
2. Il est rattaché au nominal ou verbe le plus proche, s'il n'y a pas candidat à sa rection.

On ne peut pas s'appuyer en toute confiance sur les capacités rectionnelles des verbes et des nominaux pour prédire le rattachement de tel SP qui les suit, du fait de la forte ambiguïté des prépositions P (SP→ P + N'') . Un SP introduit par P suivant un verbe ou un nominal susceptible de régir un datif, locatif, accusatif ou ablatif n'est pas nécessairement régi par ce verbe ou ce nominal.

V.3.2.7- Analyseur morpho-syntaxique

Le noyau d'analyse du français est découpé en modules principaux : morphologie et syntaxe, qui sont à leur tour répartis en sous-modules. Pour la morphologie, nous avons explicitement décrit les sous-modules, partant de la segmentation superficielle jusqu'à l'analyse flexionnelle.

Quant à l'analyse syntaxique, elle s'opère sur le texte issu de l'analyse morphologique et recouvre de sous-modules : la transcription formelle de la grammaire de réécriture des propositions (phrases, syntagmes) en automates (ATN), l'adjonction des contraintes d'accord et de rattachement (du SP) à la grammaire, ainsi que le découpage des phrases complexes en propositions simples tout en construisant leurs structures (ATN en cascades). L'analyse syntaxique proprement dite se donne dans sa fonction de construire les structures syntaxiques des différentes propositions. L'analyse syntaxique de la phrase suppose l'identification des syntagmes et la structuration hiérarchique de ces syntagmes en liaison avec leurs fonctions.

Dans le but de l'étude sur corpus que nous nous sommes fixé, l'analyse automatique des résumés textuels qui sont issus des notices en vue de l'indexation au moyen des syntagmes nominaux, comporte une série d'automates ATN (de W. Woods) développées [WOODS, 97].

La chaîne à analyser correspond à une *série d'objets* comportant chacun la forme régularisée et son profil syntaxique, lexical et flexionnel. L'automate ATN lit successivement dans la série et selon le profil syntaxique de l'objet lu, passe ou non à un nouvel état.

Théoriquement l'automate est caractérisé par les automatismes (machines) suivant(e)s :

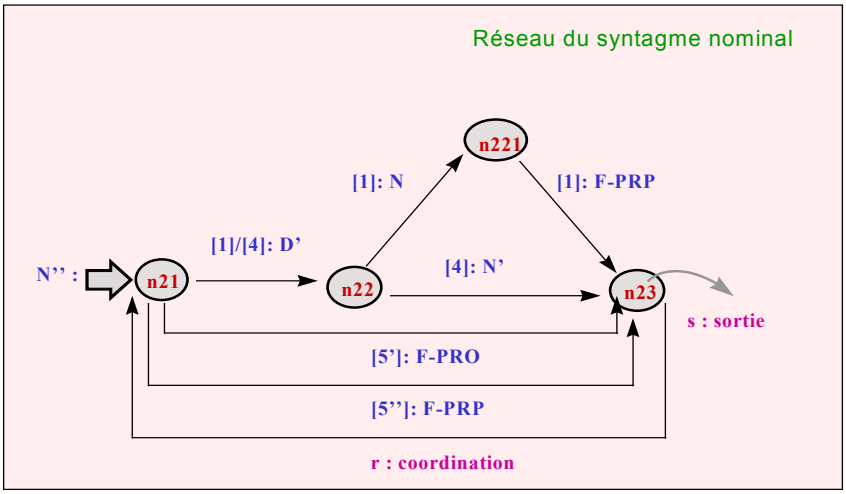
1. Un automatisme qui traite la proposition principale et débute dès l'appel du type de la proposition déclenchée ;
2. Un automatisme sous-jacent et récursif à l'automate principal et se déclenche pour traiter les subordonnées de la proposition principale ;
3. Un automatisme en cascade (CATN) permet de relier la sortie de certains automatismes comme entrée pour d'autres : ce principe d'automatisme permet la généralisation de la notion des ATN (machine type 1. ou 2.).

Nous consacrons la mise en oeuvre de ces automatismes dans le paragraphe suivant.

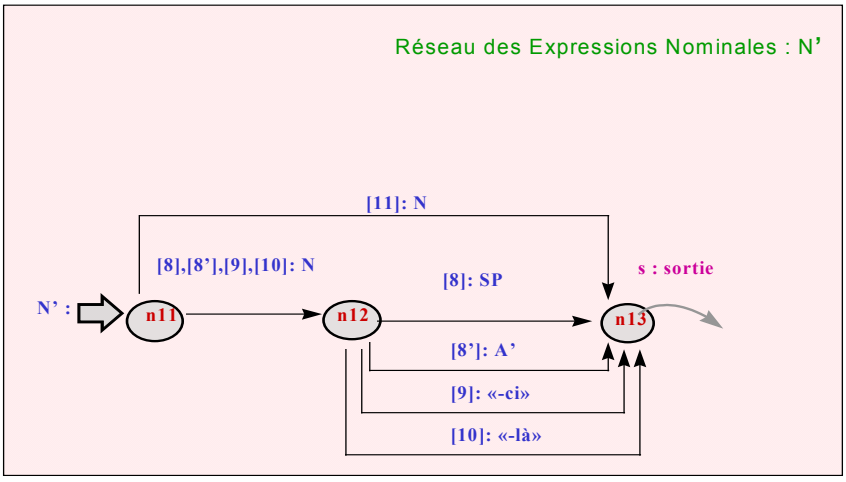
V.3.3- Mise en oeuvre des automates

V.3.3.1- Les automates du syntagme nominal

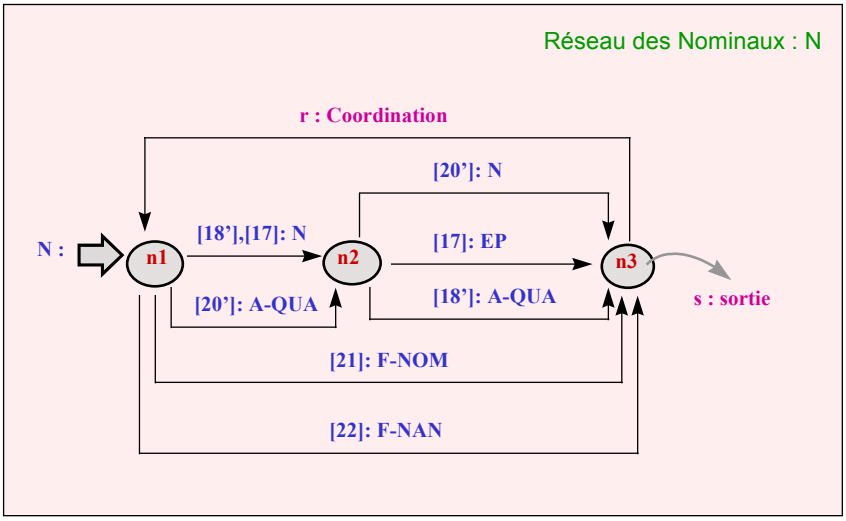
Description	N° règle	Règle
syntagmes nominaux :	1	N''→ D' + N + F-PRP
	2	N''→ D' + N'
	3	N''→ NOM-PRO
	4	N''→ NOM-PRP



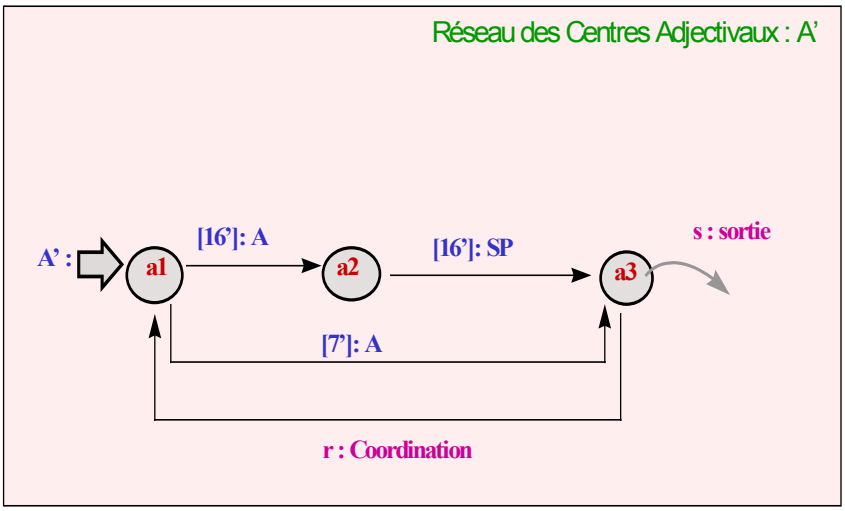
Description	N° règle	Règle
expressions nominales :	5	N'→ N + SP + (SP)
	6	N'→ N + A'
	7	N'→ N + CI
	8	N'→ N + LA
	9	N'→ N



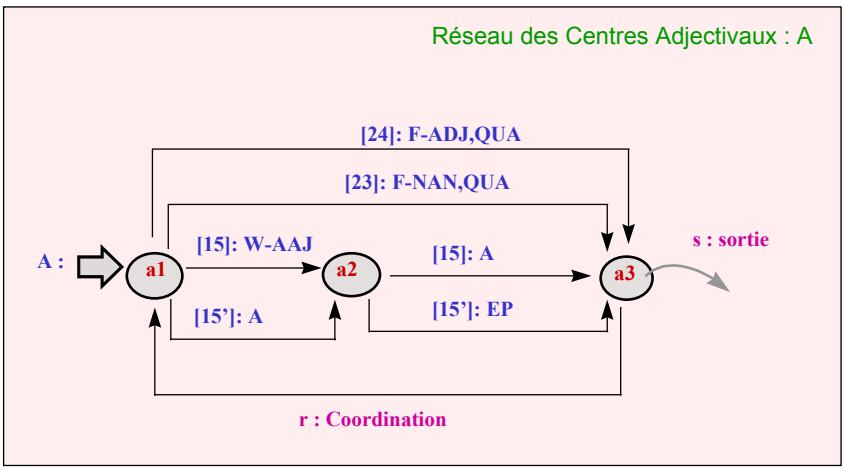
Description	N° règle	Règle
centres nominaux :	19	N→ N + EP
	20	N→ N + A(QUA)
	21	N→ A(QUA) + N



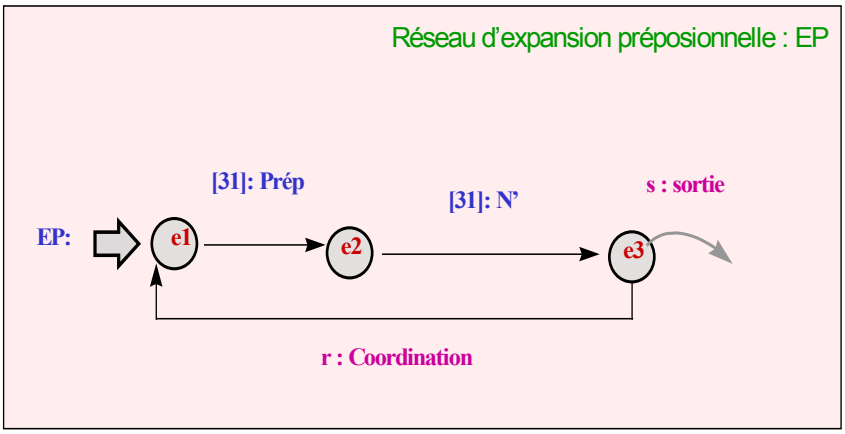
Description	N° règle	Règle
centres adjectivaux :	15	A'→ W-AAJ + A
	16	A'→ A + EP
	17	A'→ F-ADJ,REL
	18	A'→ A



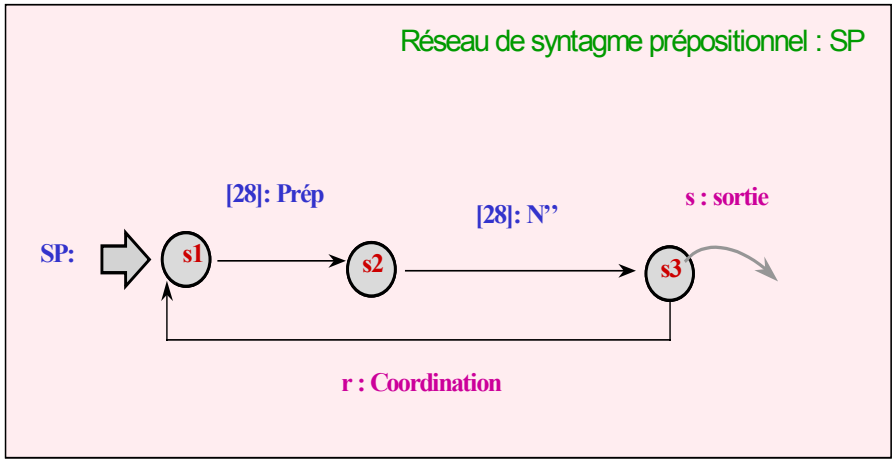
Description	N° règle	Règle
	24	A→ F-NAN,(QUA)
	25	A→ F-ADJ,(QUA)



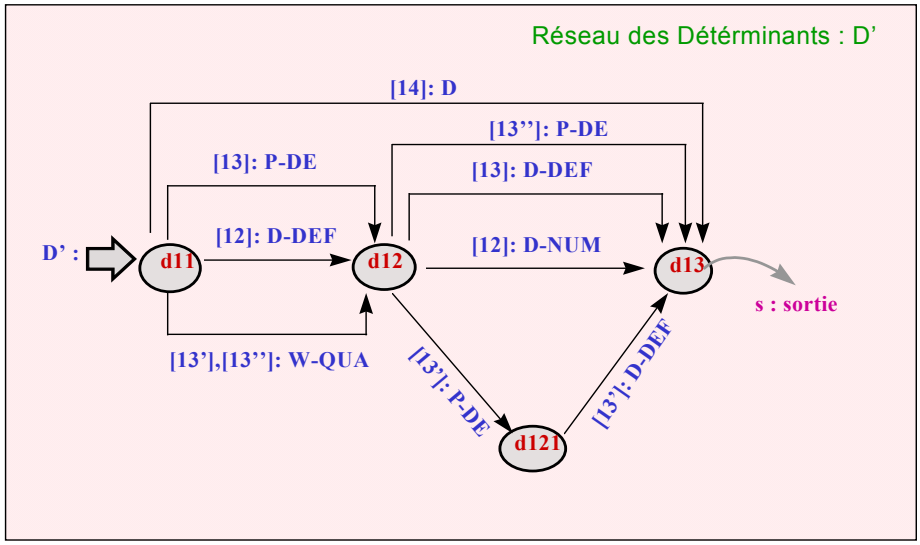
Description	N° règle	Règle
expansion prépositionnelle :	27	EP→ P + N'



Description	N° règle	Règle
syntagme prépositionnel :	26	SP→ P + N''



Description	N° règle	Règle
expressions prédéterminatives :	10	D'→ D-DEF + D-NUM
	11	D'→ P-DE + D-DEF
	12	D'→ W-QUA + P-DE + D-DEF
	13	D'→ W-QUA + P-DE
	14	D'→ D

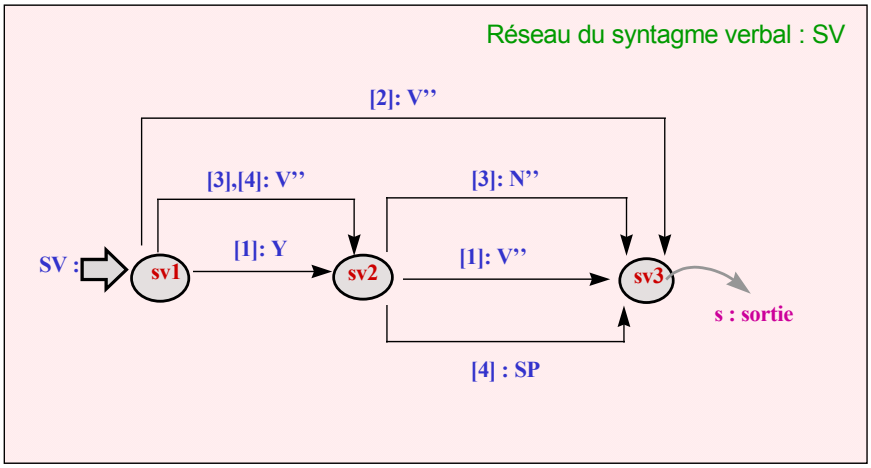


- Cas particuliers de SN

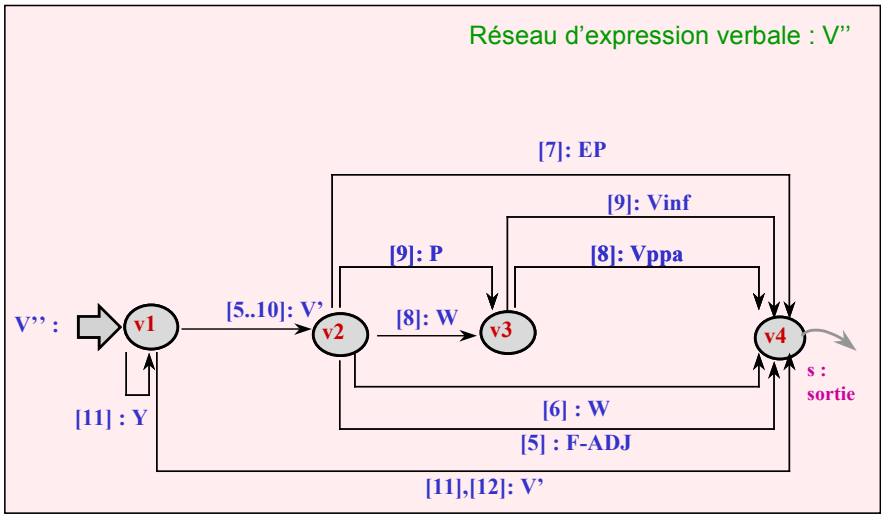
Description	N° règle	Règle
SN avec D'-zero : $SN^{\underline{z}}$	1	N'' → ∅ + N'
Syntagme date : $SN^{\underline{d}}$	2	N''→ D' + K + N' + (K)
Syntagme mesure : $SN^{\underline{m}}$	3	N''→ K + N'
Expression quantitative : K	4	K → unité-de-mesure

V.3.3.2- Les automates du syntagme verbal

Description	N° règle	Règle
Syntagme verbal :	1	SV→ Y + V"
	2	SV→ V"
	3	SV→ V" + N"
	4	SV→ V" + SP

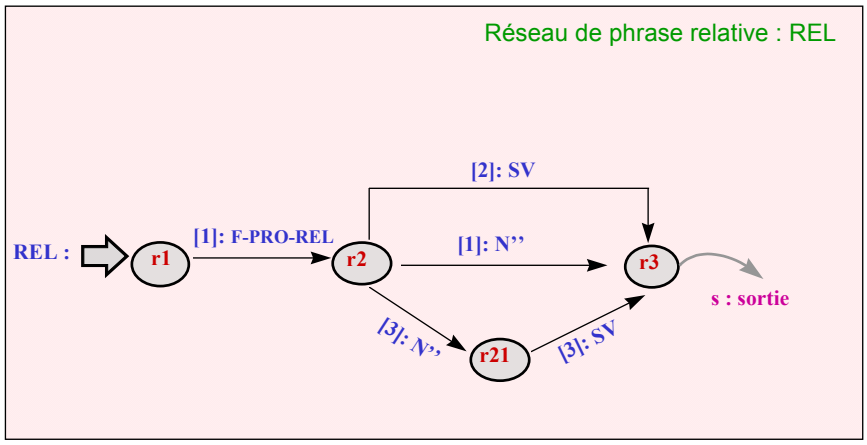


Description	N° règle	Règle
Expressions verbales :	5	V"→ V' + F-ADJ
	6	V"→ V' + W
	7	V"→ V' + EP
	8	V" → V' + W + PPAS
	9	V" → V' + P + VINF
	10	V" → V' + W + P + VINF
	11	V" → Y* + V'
	12	V" → V'
Centres verbaux :	13	V' → V + V
	14	V' → V



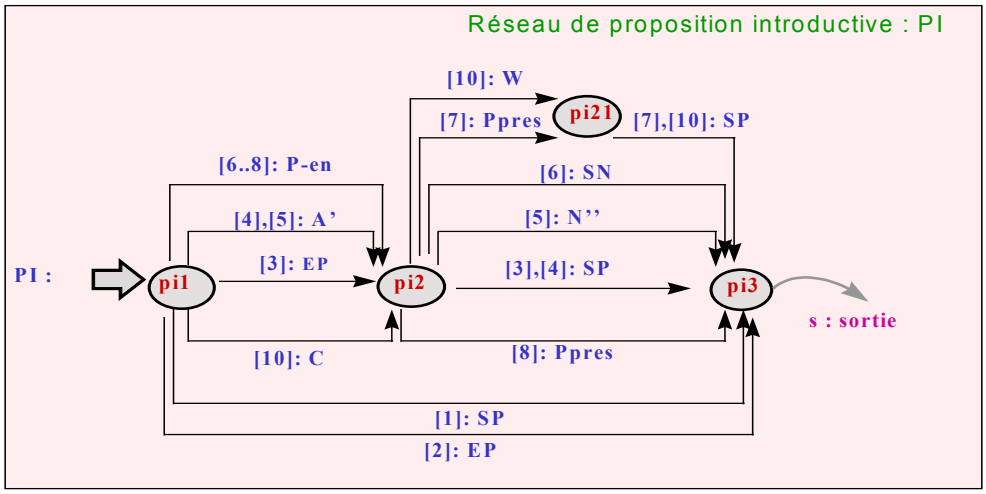
V.3.3.3- Automate de la phrase relative

Description	N° règle	Règle
Phrase relative :	1	REL → F-PRO-REL + N''
	2	REL → F-PRO-REL + SV
	3	REL → F-PRO-REL + N'' + SV



V.3.3.4- Automate de la proposition introductive de la phrase

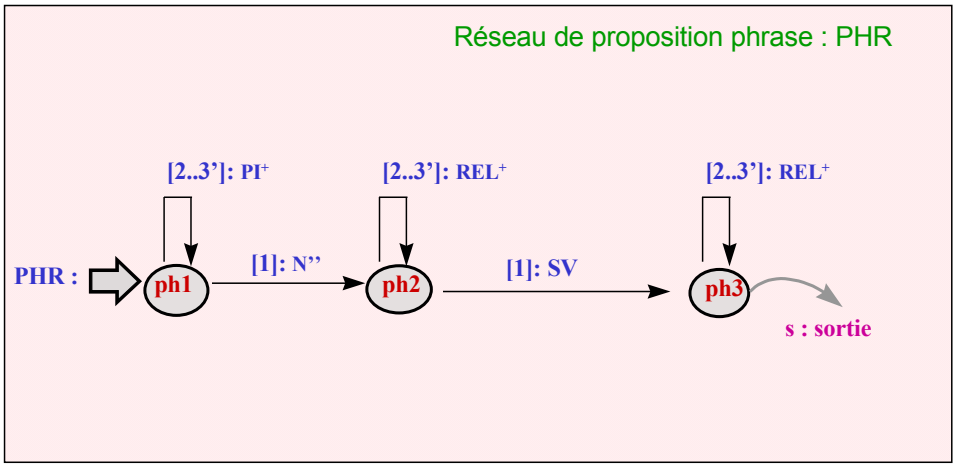
Description	N° règle	Règle
Proposition introductive :	1	PI → SP
	2	PI → EP
	3	PI → EP + SP
	4	PI → A' + SP
	5	PI → A' + N''
	6	PI → P-en + SN^{d}
	7	PI → P-en + PPRES + SP
	8	PI → P-en + PPRES
	9	PI → C
	10	PI → C + W + SP



V.3.3.5- Automate de la phrase

Description	N° règle	Règle
phrase :	1	PHR → N'' + SV
	2	PHR → PI + N'' + SV
	3	PHR → (PI) + N'' + (REL) + SV

	3’	PHR → (PI) + N’’ + SV + (REL)



V.3.4- Représentation interne engendrée par les automates

Chaque type d’automate (principal, sous-jacent ou CATN respectivement machine type 1,2,3) fonctionne selon les mêmes principes d’un ATN et consiste à vérifier des conditions et des actions au niveau des arcs de transition : – les conditions vont restreindre la possibilité de traverser sur un arc, et – les actions vont construire la structure interne de la proposition en question (phrase ou syntagme).

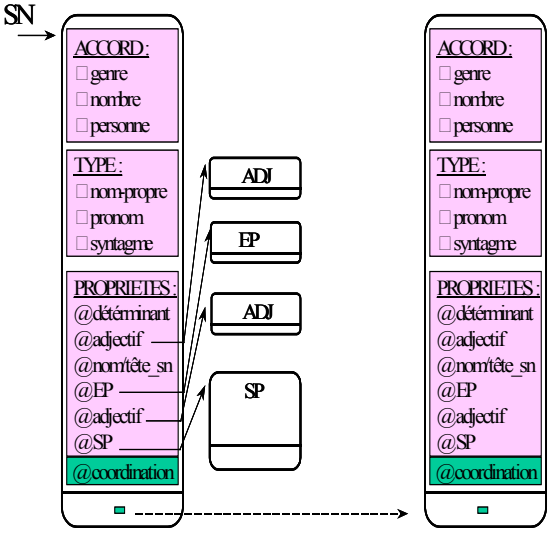
Il s’agit de spécifier les caractéristiques de chaque type d’action : la représentation interne engendrée par chaque type d’automate.

V.3.4.1- Représentation interne du syntagme nominal

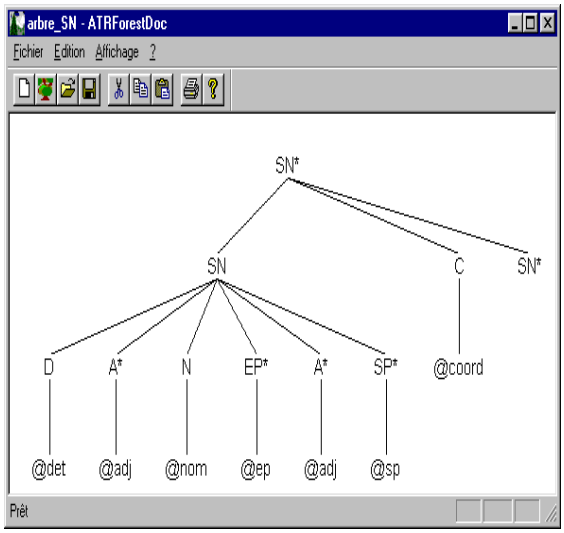
La structure interne du SN consiste initialement à créer un squelette d’objets structurés : ACCORD, TYPE et PROPRIETES. Chacun de ces objets initialement vide peut établir un lien dynamique avec des contenus simples ou complexes selon les objets auxquels ils sont dédiés.
Les contenus simples sont ceux qui se retrouvent dans ACCORD et TYPE. Les contenus complexes sont attribués à PROPRIETES. Les contenus complexes sont assimilés à des objets structurés et ayant les mêmes propriétés que l’arbre du SN.

Le remplissage des contenus s’effectuera pendant le fonctionnement des automates aussi bien que la connexion ou la modification des objets.

L’objet final obtenu lors de la validation par l’automate est un arbre décoré (arbre syntagmatique du SN).Un lien de coordination pourra s’établir avec un nouvel objet de type SN.



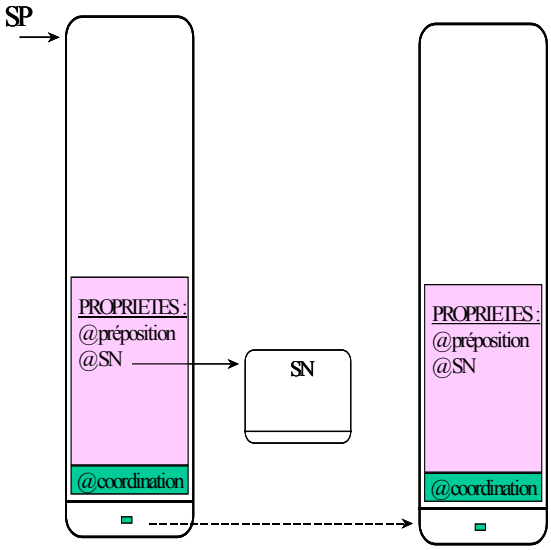
Squelette de SN



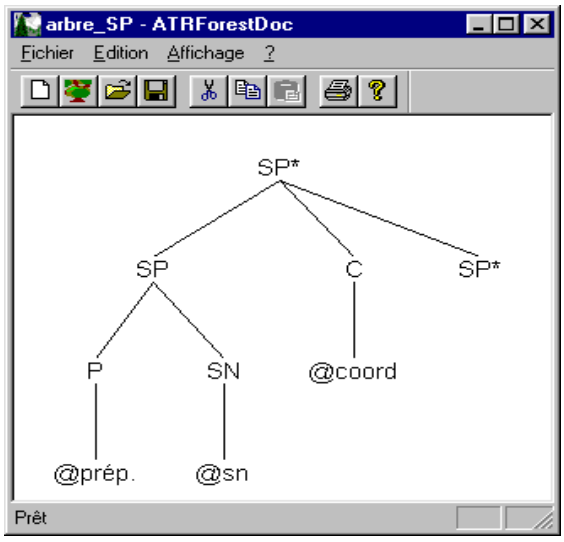
Objets liés à SN

V.3.4.2- Représentation interne du syntagme prépositionnel

La structure interne du SP consiste initialement à créer un squelette d'objet structuré : PROPRIETES. Cet objet initialement vide établira le lien avec l'objet SN. Un lien de coordination pourra s'établir avec un nouvel objet de type SP.



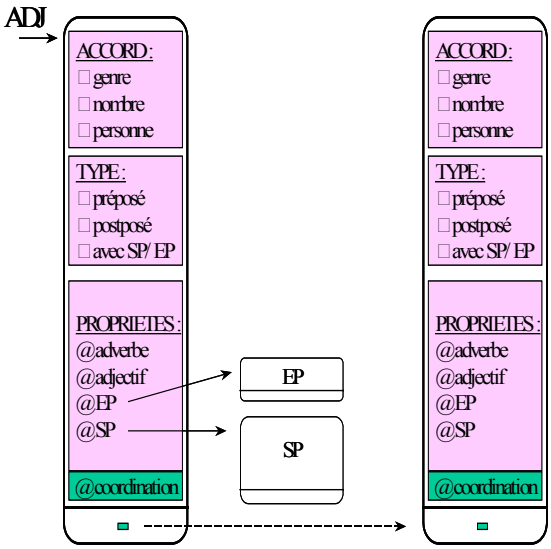
Squelette de SP



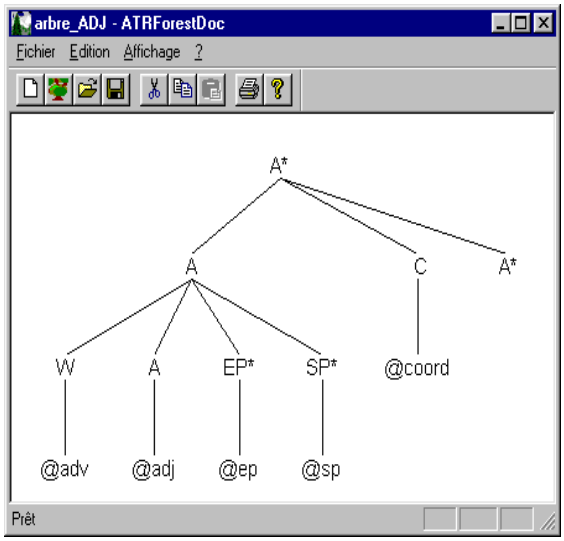
Objets liés à SP

V.3.4.3- Représentation interne du syntagme adjectival

La structure interne de ADJ (ou SA : syntagme adjectival) est identique au squelette d'un objet SN avec la différence des contenus (PROPRIETES) dans chaque objet. Un lien de coordination pourra s'établir avec un nouvel objet de type ADJ.



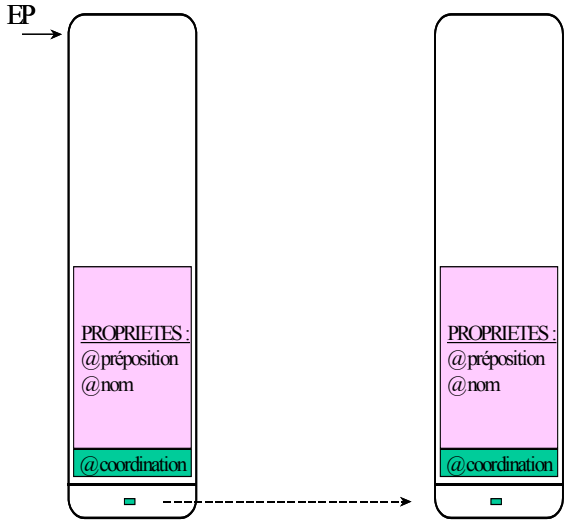
Squelette de ADJ



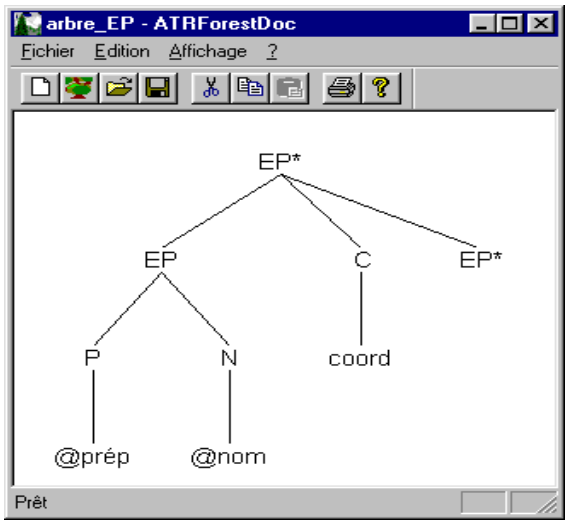
Objets liés à ADJ

V.3.4.4- Représentation interne de l'expansion prépositionnelle

La structure interne de EP est identique à l'objet SP avec la différence des noms de type simple. Un lien de coordination pourra s'établir avec un nouvel objet de type EP.



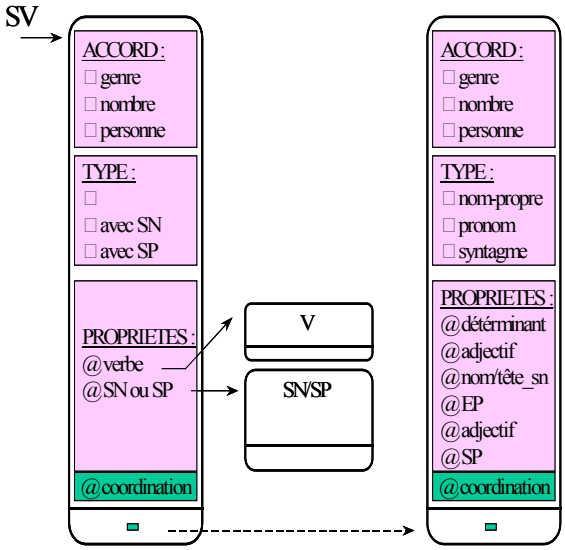
Squelette de EP



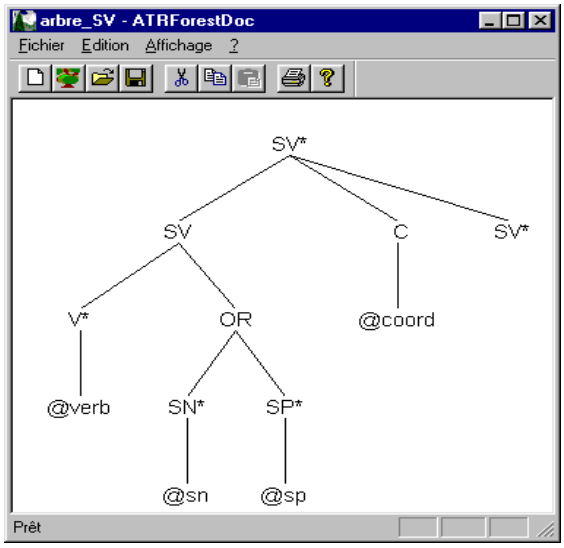
Objets liés à EP

V.3.4.5- Représentation interne du syntagme verbal

La structure interne de SV est identique au squelette d'un objet SN avec la différence des noms de contenus et leurs types dans chaque objet. Un lien de coordination pourra s'établir avec un nouvel objet de type SV.



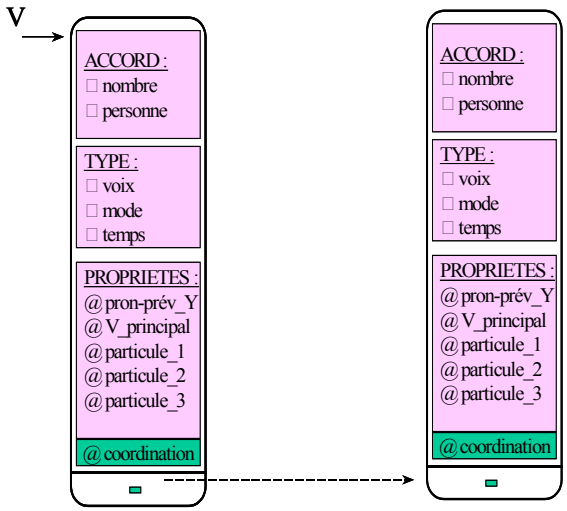
Squelette de SV



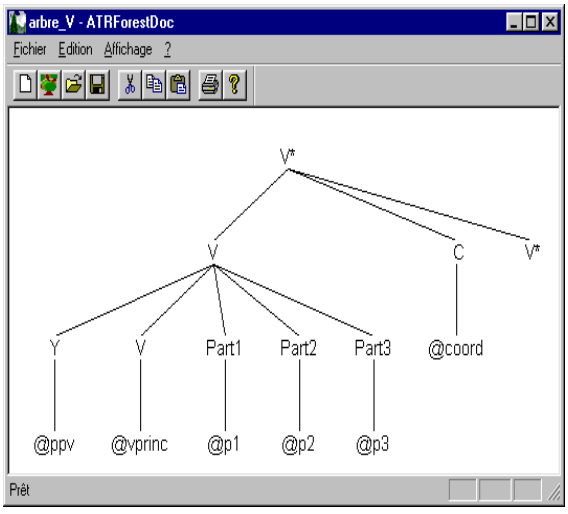
Objets liés à SV

V.3.4.6- Représentation interne du verbe

idem. que la structure interne de SV. Un lien de coordination pourra s'établir avec un nouvel objet de type SV.



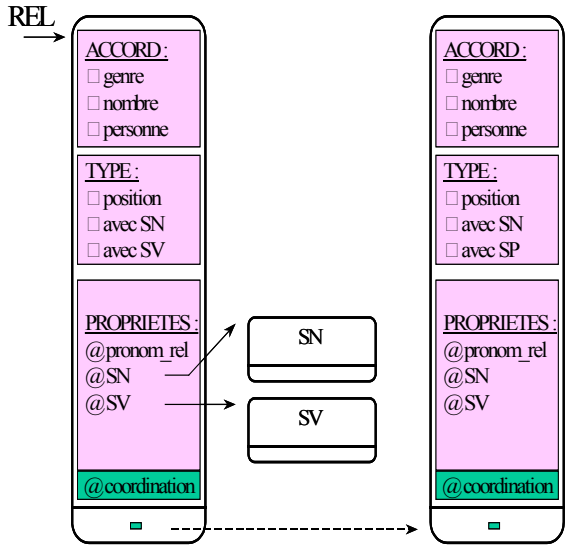
Squelette de V



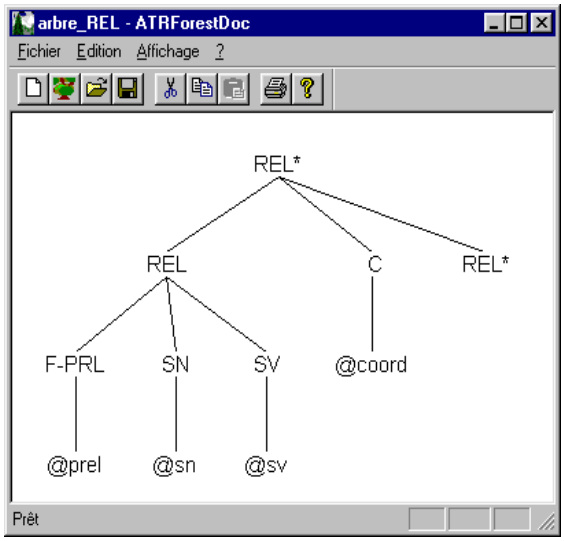
Objets liés à V

V.3.4.7- Représentation interne de la phrase relative

La structure interne de REL est de base composée d'objets complexes pouvant relier des objets de type SN et/ou SV eux mêmes de type complexe. Un lien de coordination pourra s'établir avec un nouvel objet de type REL.



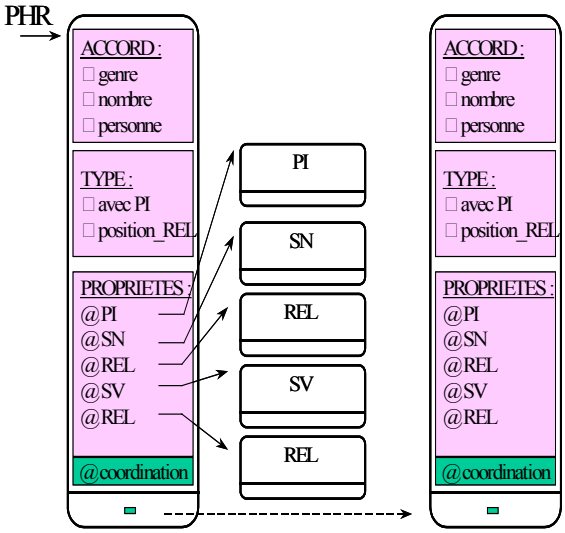
Squelette de REL



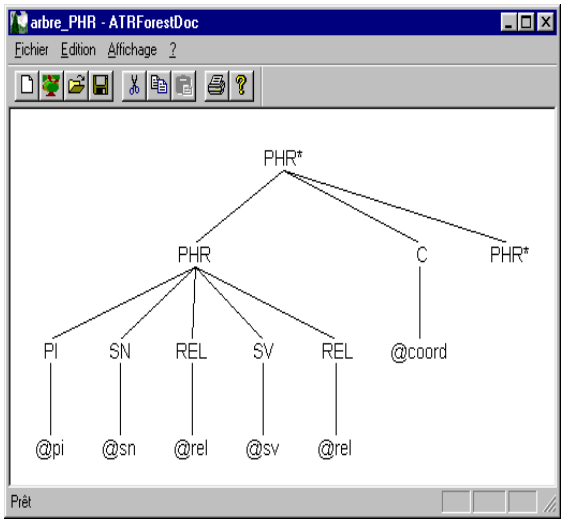
Objets liés à REL

V.3.4.8- Représentation interne de la phrase

idem. que la structure REL.



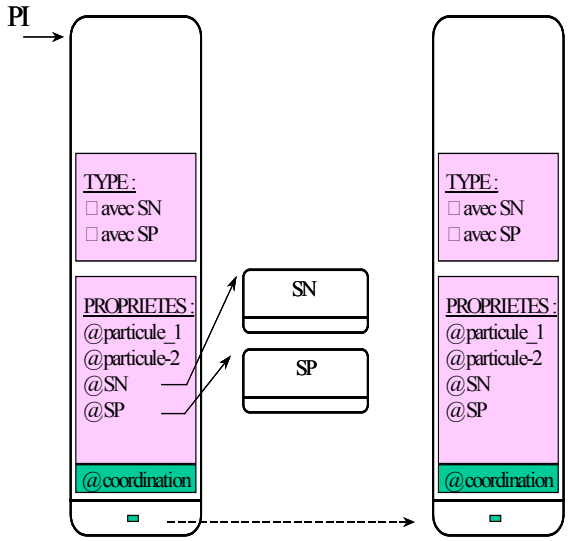
Squelette de PHR



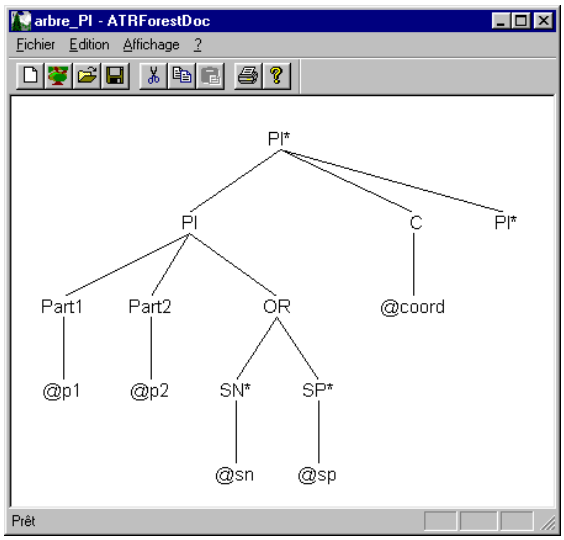
Objets liés à PHR

V.3.4.9- Représentation interne de la proposition introductive

idem. que la structure SV.



Squelette de PI



Objets liés à PI

V.3.5- Synchronisation entre le fonctionnement des automates et les structures internes engendrées

Le fonctionnement des automates est synchronisé avec les objets structurels mémorisant les constructions syntaxiques. Ces constructions syntaxiques ou structures internes sont engendrées lors de l'appel et de l'analyse par les automates (*parsing*). A chaque passage de l'automate d'un état à un autre, plusieurs opérations sont effectuées, à savoir principalement :

- Opération de lecture dans la chaîne textuelle : élément de la liste de mots
- Opération de tests de passage d'un état (i) à (i+1) : conditions sur l'arc de passage
- Opération de création d'un objet structurel : squelette de X qui est créée vide
- Opération d'activation d'un automate : automate appelé
- Opération de mise à jour d'un élément : squelette X qui reçoit un objet structurel
- Opération d'avancer dans la lecture : élément suivant de la liste

Toutes ces opérations sont inhérentes au mécanisme de l'automate pour permettre une mise en oeuvre indépendante par rapport aux types. Ci-dessous, les tableaux récapitulatifs illustrent fidèlement toutes ces opérations et permettent de transcrire l'algorithme de fonctionnement pour chaque type d'automate implémenté.

V.3.5.1- Algorithme du syntagme nominal N'' (ou SN)

Automate	Etat(i)	Arc	Condition(s)	Etat(i+1)	Action(s)
N''	n21	-	-	-	Créer(N″)
	n21	1,4	? D' ? Arc_D' vrai	n22	Activer(D') N″.déterminant← D' Avancer
	n21	5'	? F-PRO	n23	N″.pronom_tonique← F-PRO Avancer

	n21	5''	? F-PRP	n23	N″.tete_sn← F-PRP Avancer
	n22	1	? N	n221	Activer(N) N″.*← N Avancer
	n221	1	? F-PRP	n23	N″.nom_propre← F-PRP Avancer
	n22	4	? N'	n23	Activer(N') N''.*← N' Avancer

V.3.5.2- Algorithme des expressions nominales (N')

Automate	Etat(i)	Arc	Condition(s)	Etat(i+1)	Action(s)
N'	n11	-	? référence(N″)	-	-
	n11	11	? N	n13	Activer(N) N″.*← N Avancer
	n11	8	? N ? *.suivant = SP	n12	Activer(N) N″.*← N Avancer
	n11	8'	? N ? *.suivant = A'	n12	Activer(N) N″.*← N Avancer
	n11	9	? N ? *.suivant = « ci »	n12	Activer(N) N″.*← N Avancer
	n11	10	? N ? *.suivant = « là »	n12	Activer(N) N″.*← N Avancer
	n12	8	? SP	n13	Activer(SP) N″.syntagme_sp← SP Avancer
	n12	8'	? A' ? Arc_A' vrai	n13	Activer(A') N″.syntagme_adj← A' Avancer
	n12	9	? « ci »	n13	N″.tete_sn← « -ci » Avancer
	n12	10	? « là »	n13	N″.tete_sn← « -là » Avancer

V.3.5.3- Algorithme des centres nominaux (N)

Automate	Etat(i)	Arc	Condition(s)	Etat(i+1)	Action(s)
N	n1	-	? référence(N″)	-	-
	n1	21	? F-NOM(-COM)	n3	N″.tete_sn← F-NOM Avancer
	n1	22	? F-NAN	n3	N″.tete_sn← F-NAN Avancer
	n1	18'	? F-NOM ? *.suivant = A	n2	N″.tete_sn← F-NOM Avancer
	n1	17	? F-NOM ? *.suivant = EP	n2	N″.tete_sn← F-NOM Avancer

	n1	20’	? A	n2	Activer(A) N″.syntagme_adj(préposé)← A Avancer
	n2	18’	? A	n3	Activer(A) N″.syntagme_adj(postposé)← A Avancer
	n2	17	? EP	n3	Activer(EP) N″.expansion_ep← EP Avancer
	n2	20’	? F-NOM	n3	N″.tete_sn← F-NOM Avancer

V.3.5.4- Algorithme du syntagme adjectival A’’ (SA)

Automate	Etat(i)	Arc	Condition(s)	Etat(i+1)	Action(s)
A’’	a1	-	-	-	Créer(A″)
	a1	7’	? A	a3	Activer(A) A″.*← A Avancer
	a1	16’	? A ? *.suivant = SP	a2	Activer(A) A″.*← A Avancer
	a2	16’	? SP	a3	Activer(SP) A″.syntagme_sp← SP Avancer

V.3.5.5- Algorithme des centres adjectivaux (A)

Automate	Etat(i)	Arc	Condition(s)	Etat(i+1)	Action(s)
A	a1	-	? référence(A″)	-	-
	a1	24	? F-ADJ	a3	A″.adjectif← F-ADJ Avancer
	a1	23	? F-NAN	a3	A″.adjectif← F-NAN Avancer
	a1	15	? W-AAJ ? *.suivant = A	a2	A″.adverbe← W-AAJ Avancer
	a1	15’	? A ? *.suivant = EP	a2	A″.adjectif← A Avancer
	a2	15	? A	a3	A″.adjectif← A Avancer
	a2	15’	? EP	a3	Activer(EP) A″.expansion_ep← EP Avancer

V.3.5.6- Algorithme de l’expansion prépositionnelle (EP)

Automate	Etat(i)	Arc	Condition(s)	Etat(i+1)	Action(s)
EP	e1	-	-	-	Créer(EP)
	e1	31	? P	e2	EP.préposition← P Avancer
	e1	31	? N (F-NOM)	e3	EP.nom← N Avancer

V.3.5.7- Algorithme du syntagme prépositionnel (SP)

Automate	Etat(i)	Arc	Condition(s)	Etat(i+1)	Action(s)
SP	s1	-	-	-	Créer(SP)
	s1	31	? P	s2	SP.préposition← P Avancer
	s2	31	? N''	s3	Activer(N'') SP.syntagme_sn← N'' Avancer

V.3.5.8- Algorithme des expressions prédéterminatives (D'')

Automate	Etat(i)	Arc	Condition(s)	Etat(i+1)	Action(s)
D''	d11	-	? référence(N″)	-	-
	d11	14	? D	d13	D''.déterminant← D Avancer
	d11	12	? D-DEF ? *.suivant=D-NUM	d12	D''.déterminant← D-DEF Avancer
	d11	13	? P-DE	d12	D''.déterminant ← P-DE Avancer
	d11	13',13''	? W-QUA	d12	D''.déterminant ← W Avancer
	d12	12	? D-NUM	d13	D''. déterminant ← D-NUM Avancer
	d12	13	? D-DEF	d13	D''. déterminant ← D-DEF Avancer
	d12	13''	? P-DE	d13	D''. déterminant ← P-DE Avancer
	d12	13'	? P-DE	d121	D''. déterminant ← P-DE Avancer
	d121	13'	? D-DEF	d13	D''. déterminant ← D-DEF Avancer

V.3.5.9- Algorithme du syntagme verbal (SV)

Automate	Etat(i)	Arc	Condition(s)	Etat(i+1)	Action(s)
SV	sv1	-	-	-	Créer(SV)
	sv1	2	? V'	sv3	Activer(V') SV.verbe← V' Avancer
	sv1	1	? Y	sv2	SV.pronom_préverbal← Y Avancer
	sv1	3,4	? V'	sv2	Activer(V') SV.verbe← V' Avancer
	sv2	1	? V'	sv3	Activer(V'') SV.verbe← V' Avancer
	sv2	3	? N'' ? Arc_D' vrai	sv3	Activer(N'') SV.syntagme_sn← N'' Avancer
	sv2	4	? SP	sv3	Activer(SP) SV. syntagme_sp← SP Avancer

V.3.5.10- Algorithme des expressions verbales (V')

Automate	Etat(i)	Arc	Condition(s)	Etat(i+1)	Action(s)
V'	v1	-	-	-	Créer(V')
	v1	5..12	? Y	v1	V'. pronom_préverbal ← Y Avancer
	v1	11,12	? V	v4	V'.verbe← V Avancer
	v1	5..10	? V ? *.suivant={P,W,A}	v2	V'.verbe← V Avancer
	v2	5	? F-ADJ	v4	V'.post_verbal← F-ADJ Avancer
	v2	6	? W	v4	V'.post_verbal← W Avancer
	v2	7	? EP	v4	Activer(EP) V'.post_verbal← EP Avancer
	v2	8	? W ? *.suivant= Vppa	v3	V'.particule_verbale← W Avancer
	v2	9	? P	v3	V'.particule_verbale← P Avancer
	v3	8	? Vppa	v4	V'.particule_verbale← Vppa Avancer
	v3	9	? Vinf	v4	V'.particule_verbale← Vinf Avancer

V.3.5.11- Algorithme de la phrase relative (REL)

Automate	Etat(i)	Arc	Condition(s)	Etat(i+1)	Action(s)
REL	r1	-	-	-	Créer(REL)
	r1	1..3	? F-PRO-REL	r2	REL.pronom_rel← F-PRO-REL Avancer
	r2	1	? N'' ? Arc_D' vrai	r3	Activer(N'') REL.syntagme_sn ← N'' Avancer
	r2	2	? SV	r2	Activer(SV) REL.syntagme_sv ← SV Avancer
	r2	3	? N'' ? Arc_D' vrai	r21	Activer(N'') REL.syntagme_sn ← N'' Avancer
	r21	3	? SV	r3	Activer(SV) REL.syntagme_sv ← SV Avancer

V.3.5.12- Algorithme de la proposition introductive (PI)

Automate	Etat(i)	Arc	Condition(s)	Etat(i+1)	Action(s)
PI	pi1	-	-	-	Créer(PI)
	pi1	1	? SP	pi3	Activer(SP) PI.syntagme_sp ← SP Avancer
	pi1	2	? EP	pi3	Activer(EP)

					PI.expansion_ep← EP Avancer
	pi1	3	? EP ? *.suivant= SP	pi2	Activer(EP) PI.expansion_ep← EP Avancer
	pi1	4, 5	? A' ? *.suivant= SP : 4 ? *.suivant= N'': 5	pi2	Activer(A') PI.syntagme_adj ← A' Avancer
	pi1	6..8	? P-en ? *.suivant= SN^{d} : 6 ? *.suivant=Ppres:7,8	pi2	PI.particule← P-en Avancer
	pi1	10	? C ? *.suivant= W	pi2	PI.particule← C Avancer
	pi2	3,4	? SP	pi3	Activer(SP) PI.syntagme_sp ← SP Avancer
	pi2	5	? N''	pi3	Activer(N'') PI.syntagme_sn← N'' Avancer
	pi2	6	? SN^{d}	pi3	Activer(N'') PI.syntagme_sn← SN^{d} Avancer
	pi2	8	? Ppres	pi3	PI.particule← Ppres Avancer
	pi2	7	? Ppres ? *.suivant= SP	pi21	PI.particule← Ppres Avancer
	pi2	10	? W	pi21	PI.particule← W Avancer
	pi21	7,10	? SP	pi3	Activer(SP) PI. syntagme_sp← SP Avancer

V.3.5.13- Algorithme de la phrase (PHR)

Automate	Etat(i)	Arc	Condition(s)	Etat(i+1)	Action(s)
PHR	ph1	-	-	-	Créer(PHR)
	ph1	2..3'	? PI	ph1	Activer(SP) PHR.pi← PI Avancer
	ph1	1	? N''	ph2	Activer(N'') PHR.syntagme_sn← N'' Avancer
	ph2	2..3'	? REL	ph2	Activer(REL) PHR.relative← REL Avancer
	ph2	1	? SV	ph3	Activer(SV) PHR.syntagme_sv← SV Avancer
	ph3	2..3'	? REL	ph3	Activer(REL) PHR.relative← REL Avancer

Cependant, nous n'avons pas décrit, dans nos travaux, tous les types d'opérations algorithmiques des automates, comme le calcul des accords et la communication entre les automates.

V.4- Limites du modèle syntaxique implémenté

L'enjeu et les contraintes qui pèsent sur le choix du modèle syntaxique et son analyseur consistent à trouver les structures syntaxiques engendrées par la grammaire sur un segment de phrase ou une phrase entière. Ce segment correspond à une ou plusieurs propositions au sens défini par le modèle classificatoire.
La grammaire proposée et implémentée est ambiguë, car elle engendre souvent sur une même séquence plus d'une structure syntaxique. Les causes des solutions multiples sont :

- soit la séquence analysée admet plusieurs analyses possibles (structures) attestées par la langue,
- soit la grammaire de réécriture des structures syntaxiques engendre plusieurs solutions pour une même séquence analysée. Certaines étant attestées par la langue, d'autres dûes à des solutions parasites (non attestées).

La stratégie d'analyse devrait permettre d'améliorer la justesse des solutions. Les données à partir desquelles la stratégie s'opère sont des informations linguistiques portées par la séquence à analyser. Un superviseur de type combinatoire rentre dans cette stratégie de manière à ajuster les solutions attestées par la langue sur l'ensemble des solutions possibles. Son rôle consiste à piloter l'analyseur vers les règles qui engendrent les bonnes structures, réduire les solutions possibles en éliminant les cas parasites, et modifier si nécessaire l'ordre de surface pour permettre l'analyse à l'aide de la grammaire de la proposition.

Face à ses limites, d'une part, associées au modèle, et d'autre part, à la stratégie de sa mise en oeuvre, l'apport des solutions n'est pas définitif à chacun de ces problèmes. Simplement, les propositions de solutions employées sont étroitement liées aux difficultés observées et les limitations sur des essais d'analyse sur le corpus. Il reste d'autres problèmes qui dépassent le cadre fixé de notre travail comme le traitement complexe de la coordination, le traitement de la rection verbale, etc.

V.5- Conclusion

« il s'agit avant tout d'un problème linguistique » tel est l'expression retenue par A. Berrendonner et M. Le Guern face aux problèmes liées à l'analyse morpho-syntaxique et à l'interprétation d'un énoncé écrit pour une représentation logique.

Dans le cadre du travail présenté, le problème d'indexation automatique des énoncés écrits est intrinsèquement lié à cette problématique d'ordre linguistique.

Le passage de l'énoncé en langue naturelle à une expression logique (structuration) du formalisme recouvre deux étapes successives.
La première qui s'arrête à l'issue de l'analyse morpho-syntaxique fait l'objet de ce chapitre.
Elle décrit une partie de notre projet informatique.
La seconde concerne la réutilisation de cet objet et son exploitation pour l'indexation automatique versus une application de recherche d'information basée sur les syntagmes nominaux.

Cet analyseur, écrit en C et C++ (MS.Visual C++), est orienté objet.

Dans cette approche, les unités élémentaires avec lesquelles le programme travaille sont des objets. Ceux-ci sont des mots auxquels sont attachées une classe (ex.: nom) et une catégorie (ex.: sujet).

En fonction de leurs caractéristiques, les objets envoient des messages à d'autres objets. Ceux-ci effectuent certaines actions. Par exemple, un certain type de nominaux va envoyer des messages vers un certain type de sujet ou de complément, et attendre des réponses.

Les actions constituent la construction progressive de l'arbre décoré ou de structures objets. Un avantage de l'écriture de l'analyseur (l'adaptation du formalisme ATN aux structures Objets) est sa grande flexibilité quant au formalisme choisi et aux actions à effectuer suite à un message.

La grammaire implémentée utilise des algorithmes d'analyse des structures en deux étapes :

- un analyseur qui énumère les structures possibles, en utilisant les règles combinatoires à partir des entrées syntaxiques du formalisme ATN ;
- puis l'analyseur résout les contraintes qui ne sont pas prises en compte dans la grammaire de réécriture des syntagmes nominaux.

Divers développements ont favorisé l'utilisation du formalisme ATN :

- création d'automates ATN ;
- connexion des ATN : les automates complexes ou en cascade (CATN) ;
- création d'algorithmes itératifs permettant la prise en charge dynamique des fonctionnement des automates ;
- outil d'extraction d'information : les syntagmes nominaux et leur emboîtement ;
- création automatique d'une base d'information (SN) sur les textes analysés et d'une base sur les requêtes à analyser.

L'environnement basé sur le formalisme des grammaires ATN / CATN de W. Woods, dans le but de préparer le développement de diverses applications en TALN [BATES, 83], [WOODS, 98b], telles l'acquisition de connaissance, la traduction automatique, l'extraction d'information, etc., catalyse une certaine ambition que comporte notre projet face à :

- l'organisation d'une base documentaire sur le multimédia par la manipulation des SN dans les résumés de contenu, et
- l'organisation des connaissances : indexation et recherche d'information dans les bases d'informations textuelles.

CHAPITRE VI :

Regroupement conceptuel pour l'organisation de connaissances autour du SN :

Classification et Recherche d'information

VI.1- Introduction

De nos jours, La recherche d'information est devenue un domaine en plein essor. Le développement des bases de données tout public et les sources d'information sur Internet ont accéléré ce processus. De nombreux modèles ont été construits sur la base du *paradigme de requêtes,* c'est à dire un mode dans lequel, l'utilisateur formule une requête en exprimant ce qu'il tente de retrouver. Le système de recherche d'information procède alors par l'invocation d'un processus pour satisfaire cette requête.

Une autre approche consistant sur le *paradigme organisation/navigation* se base sur l'hypothèse selon laquelle « les hommes préfèrent les tâches de reconnaissance plutôt que celles de description ». Selon cette hypothèse, l'idée de base consiste à organiser (ou classifier) l'information et à laisser l'utilisateur naviguer dans cette base structurée d'informations.

Il est clair que dans ce dernier paradigme organisation/navigation, la classification d'information est construite dans l'optique de structurer les connaissances pour mieux les retrouver. Dans d'autres cas de classification automatique, le but est de construire des techniques d'apprentissage symbolique pour la recherche d'information : arbre de concepts servant de support à la recherche.

Le but dans cette partie est de proposer une approche de construction automatique de classifications conceptuelles qui prenne en compte les sorties de l'analyseur morpho-syntaxique. Une telle approche montre ses intérêts dans la construction de classifications en vue d'organiser les connaissances autour du syntagme nominal. Le système de regroupement conceptuel permet ainsi de construire, de manière interactive, les différentes classifications à partir de descriptions d'objets structurés : une telle propriété est classée comme un savoir sur les objets considérés.

VI.2- La classification : Enjeux épistémologiques

La construction de classifications est aujourd'hui une préoccupation première dans de nombreux domaines tels que les sciences humaines, les sciences médicales, les sciences cognitives et les sciences de l'information et de la communication.
Or, Aristote est considéré comme le pionnier à avoir proposer des classifications dans le domaine de la biologie et d'en avoir fait un outil heuristique. Le souci de construire des

classes n'a pas été initialement de découvrir un ordre unique ou un ordre strictement hiérarchisé. La motivation originelle est de chercher dans −une préoccupation économique une formulation concise des informations et −une visée rationaliste permettant d'établir une corrélation significative entre les objets du monde et leurs propriétés. A la même époque, Platon s'est intéressé à la classification des êtres vivants, à celle des régimes politiques et aux multiples formes de la pensée humaine.
En revanche, les travaux issus du *courant classificatoire* au XIIème siècle portent principalement sur la classification du monde vivant : constitution de l'histoire naturelle.

Ainsi, les classifications constituent une forme essentielle de la −*connaissance* humaine, une des façons principales de −*mémoriser* les objets, les situations et les expériences vécues. Sans un procédé de classification et sans classes, le monde des entités et des objets reste individuel isolé et il serait un monde « plat » [BOURNAUD, 2000a].

En particulier, les classifications apparaissent fondamentales dans le domaine du langage. L'emploi des mots dans la langue dépend d'une classification [BERRENDONNER, 90], d'une détermination préalable de catégories. La construction et l'interprétation des phrases intelligibles supposent un savoir partagé (les catégories de la langue) et une compréhension des énoncés (consensus sur leur signification) [ZINGLÉ, 97]. Cette appréhension intellectuelle exprime ou opère en soi une structure et une sémantique. Une certaine classification se caractérise par une identification, une généralisation et une association des objets du monde.

> « L'importance des classifications dans les sciences se reflète dans la grande variété des domaines où tant leur nature que leur construction ont fait, et font encore, l'objet de recherche. » *[BOURNAUD, 96]*.

VI.2.1- Quelques définitions préalables sur la classification

Dans tout système qu'il soit humain, physique, biologique ou économique, l'apprentissage est une activité essentielle à la base de toute évolution [ZIGHED, 2000]. Dans la littérature scientifique, nous rencontrons divers point de vue et des compréhensions parfois éloignées sur la notion d'apprentissage. Une proposition faite par D. MARR (1981), et reprise par S. BOUCHERON (1992), qui distinguent une classification de l'apprentissage en trois niveaux [BOURNAUD, 96], [BOURNAUD, 2000b-c].

Dans le −premier niveau se déploient les neurosciences qui tentent d'investir sur ce qu'on appelle les *fonctions supérieures du cerveau*, par exemple pour rendre compte de la faculté du langage en termes cellulaires.

Dans le −second niveau, les sciences cognitives tentent de s'éloigner du modèle animal pour s'intéresser aux conditions et possibilités d'apprentissage ayant pour postulat de base que la pensée peut être assimilée à un calcul [WINOGRAD, 83], [DIENG, 94], [FERRET, 2000]. Ainsi, il s'agit d'explorer scientifiquement l'apprentissage sous l'angle *comportemental et/ou fonctionnel*.

Enfin, les approches calculatoires considèrent l'apprentissage comme un calcul [HAASE, 96], et essayent de l'étudier de façon rigoureuse comme phénomène propre. Par exemple, l'apprentissage en informatique théorique se propose d'étudier la complexité et postule *qu'apprendre c'est converger* [WERMTER, 92-95] [KODRATOFF, 96].

Dans ce cadre classificatoire de l'apprentissage, nous nous situons au second niveau. Nous considérons les aspects fonctionnels et algorithmiques de l'apprentissage pour chercher un cadre conceptuel [PELEATON, 2000], [RAJMAN, 97] de l'organisation des connaissances autour du syntagme nominal comme unité logique, autonome, et dotée de sens.

Les notions de regroupement et de concept sont des notions clés qui permettent de situer le rôle central de la classification par définition. Nous commençons par rappeler les acceptions inhérentes à ces définitions. Une part importante de ces définitions est énoncée dans les travaux de [BOURNAUD, 96-97-98a], [FREGE, 71], [WITTGENSTEIN, 53] :

- **<u>Définition 1 :</u> Regroupement (en anglais : cluster)**

  Un *regroupement* est un ensemble d'individus, abstraits ou concrets.

  On appelle *espèce,* un regroupement d'individus de même nature. La notion d'espèce désigne un ensemble d'individus sans que tous les individus soient nécessairement connus.

  On appelle *genre* un regroupement d'espèces.

  Ce vocabulaire générique, introduit par Aristote, est utilisé par Linné de manière systématique pour désigner des niveaux taxinomiques de la classification des êtres vivants.

- **<u>Définition 2 :</u> Concept**

  Un *concept* est une fonction définie sur un domaine de référence, qui prend ses valeurs dans le domaine {vrai, faux}.

  Cette fonction discrimine les individus auxquels s'applique le concept – la fonction prend la valeur vraie et les individus, appelés instances du concept (ou référents), sont dits recouverts par le concept – de ceux auxquels il ne s'applique pas. L'extension d'un concept est l'ensemble des instances de ce concept.

  Remarquons qu'il n'est pas toujours possible d'énumérer toutes les instances d'un concept. Par exemple, le cas de concepts infinis de nombre impair, ou encore la description de concepts issus de notre imagination, etc.

  C'est un fait que chacun observe lorsqu'il se réfère à une définition dans un dictionnaire, les définitions des catégories lexicales sont des définitions intensionnelles.

- **<u>Définition 3 :</u> Description d'un concept**

  En intelligence artificielle, on distingue deux manières principales pour décrire ou définir un concept.

  – La première manière consiste à décrire un concept sous la forme d'un ensemble de *conditions nécessaires et suffisantes* (CNS). Pour décider de l'appartenance d'un individu à l'extension d'un concept, il suffit de s'assurer qu'il vérifie les CNS qui décrivent ce concept. Par exemple, pour décider si un nombre x est un nombre pair, il suffit de vérifier que le reste de l'opération x/2 est nul. Réciproquement, tout individu appartenant à l'extension du concept vérifie les CNS.

  Pour pouvoir induire une description d'un concept sous la forme de CNS il faut disposer de la totalité du domaine de référence du concept (*cf*. Définition 2). Comme cela n'est pas toujours possible, une restriction sur un ensemble de *conditions suffisantes* suffit à la description : on parle dans ce cas de définition *heuristique* du concept.

  – La seconde manière de décrire un concept est dûe à Wittgenstein (1953) qui a remis en cause la description de concept sous forme de CNS, en énonçant le principe de la

« ressemblance familiale ». Selon cette théorie, un concept n'est autre chose qu'un élément typique, auquel est associé une certaine variabilité. En exemple, le pouvoir de décider de l'appartenance d'un individu à un concept. Si l'on représente les concepts avec un certain nombre d'attributs, cette décision se prend en considérant la globalité de l'individu que l'on confronte à la globalité du concept. Si la correspondance est conforme ou « suffisante » et « sans être totale », alors l'objet sera reconnu comme une instance du concept. Cette notion de *typicalité* constitue la principale nouveauté de l'approche.

– Une approche récente, dite « *Exemplariste* », remet en cause l'interprétation « *Prototypiste*». Elle consiste à décrire un concept par des instances représentatives : des individus qui servent à délimiter le concept. Il ne s'agit pas d'une variation sur le thème de prototype, mais d'une approche basée sur une similarité conceptuelle des instances : un concept est défini par un certain nombre d'exemplaires (instances du concept) et par une théorie.

♦ **<u>Définition 4 :</u> Caractériser un regroupement**

Caractériser un regroupement consiste à donner une description dont l'extension couvre ce regroupement.

Par exemple, considérons le regroupement $\Re = \{2,8,64\}$ et le domaine de références des chiffres. Caractériser ce regroupement revient à donner une description qui contient $\Re$ dans son extension.

Soit la CNS $= \{x, n \in \mathrm{N} / x = n^2\}$, cette description contient $\Re$ dans son extension, c'est donc une caractérisation de $\Re$. L'intérêt de recourir à la caractérisation d'un regroupement est qu'elle permet, entre autre, de définir une relation de généralité entre regroupements qui correspond à la *relation d'inclusion de leurs extensions* ou *généralité extensionnelle entre concepts*.

♦ **<u>Définition 5 :</u> Généralité entre concepts**

Soit deux concepts C_a et C_b.

Le concept C_a est dit *plus général* que le concept C_b si l'extension de C_a contient l'extension de C_b. on dit encore que C_b est *plus spécifique* que C_a, ou que C_a subsume C_b de manière extensionnelle.

En apprentissage symbolique automatique, d'autres définitions sont plus opérationnelles dans la généralité entre concepts. Celles-ci se basent sur la description des concepts et non plus sur leurs extensions. Parmi les relations de subsomption qui ont été définies en apprentissage symbolique, citons à titre d'exemple la θ-subsomption [PLOTKIN, 70], l'i-implication [GANASCIA, 87], l'implication logique [NIBLET, 88], et la co-subsomption [NAPOLI, 92-94]

On pourra étendre le principe de la généralité vers la notion de généralité entre regroupement (cf. Définition 4 et Définition 5):

♦ **<u>Définition 5bis :</u> Généralité entre regroupements**

Un regroupement $\Re_a$ est dit plus général que $\Re_b$ si tous les éléments de $\Re_b$ sont aussi dans $\Re_a$ ($\Re_b$ est inclus dans $\Re_a$).

Etant donné un ensemble de concepts, chercher à les organiser selon une relation de généralité entre concepts (*cf.* Définition 5), revient précisément à constituer une

classification. Il était indispensable, de préciser les définitions antérieures pour assurer une meilleure compréhension à la problématique liée à cette partie de notre recherche.

VI.2.2- Structure d'une classification

Par souci de clarté, le terme classification est ambiguë car il désigne à la fois un processus et des résultas liés à ce processus. Pour établir cette distinction, nous employons le terme « processus (de construction) de classification », et le *résultat* de classification par « classification ».

- **<u>Définition 6 :</u> Classification**

  Une *classification* est un ensemble de concepts organisés selon une relation de généralité entre concepts.

  Une classe est un regroupement quelconque d'une classification.

- **<u>Définition 7 :</u> Identification - Détermination**

  Etant donné une classification, l'*identification* consiste à associer à un individu une ou plusieurs classes auxquelles il appartient.

- **<u>Définition 8 :</u> Taxinomie (**en anglais : Taxonomy**)**

  Une taxinomie est un système par lequel des classes sont reliées les unes les autres par inclusion de classes.

Plus une classe est inclusive dans une taxinomie, plus élevé est le niveau d'abstraction de cette classe.

Dans une taxinomie, les classes sont appelées « taxons » et les niveaux sont appelés « niveaux taxinomiques ».

Exemple :

La classification des plantes proposée par Linné est l'exemple le plus connu de taxinomie : classe → ordre → famille → genre → espèce → variété → race.

Remarques :

Une classification peut ne pas avoir une structure d'arbre. Cela s'explique par le fait que les classes d'un même niveau d'abstraction ne sont pas exclusives ou lorsqu'une d'elles est incluse dans plus d'une autre. Il s'agira dans ce cas des structures classificatoires en pyramide et des treillis (*Fig. VI.2.2*).

Nous laissons le lecteur de retrouver en détail ces types de constructions classificatoires en arbre, en pyramide et treillis dans les travaux de [ZIGHED, 2000], [DIDAY, 2000], [BERTRAND, 2001] et [BOURNAUD, 96-98b].

Fondamentalement, ces constructions se rapportent à la notion de graphe qui est une structure combinatoire. Cette dernière permet de représenter de nombreuses situations rencontrées dans des applications qui font intervenir les mathématiques discrètes et nécessitant une solution informatique.

Rappelons que la *classification pyramidale* est une technique de classification automatique permettant, à partir d'un ensemble fini d'objets et un indice d'agrégation, d'organiser ces objets en une structure de synthèse appelée pyramide.

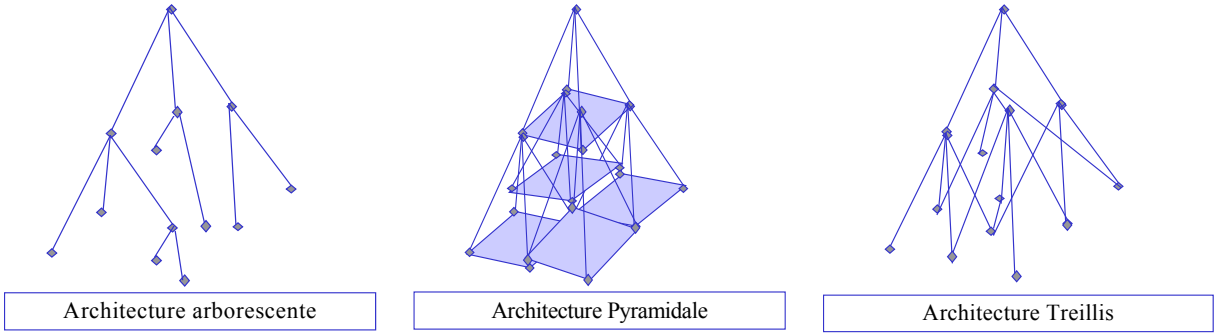


Fig. VI.2.2. Différentes structures de classifications automatiques.

VI.2.3- Processus de construction de classifications

Construire des classifications consiste, d'une part, à regrouper des individus, et d'autre part, à donner des caractéristiques de ces regroupements. La base du raisonnement qui justifie la construction de classifications n'est donc pas purement logique mais comme une forme d'apprentissage inductif (raisonnement inductif). Vis à vis de tel ou tel critère, des individus ayant certaines similarités se ressemblent et forment une classe. Cette aptitude dans le processus laisse deviner toute la complexité et la difficulté de sa mise en oeuvre.

Illustration :

Une représentation d'une classification des concepts C1, C2 et C3 dont les descriptions respectives sont les conjonctions de caractéristiques :

{d1∧d2∧d3∧d4}, {d1∧d2∧d3∧d5}, {d1∧d6}, et des extensions respectives {i1,i2,i3}, {i4,i5}, {i6, i7,i8}.

Intuitivement, le processus de classification (*Fig.VI.2.3.*) consiste à regrouper les individus ($i_{\alpha,\ \alpha=1..8}$) qui se ressemblent en classe ($C_{\beta,\ \beta=1..3}$) : c'est-à-dire qui ont certaines similarités vis à vis de tel ou tel critère ($d_{\lambda,\ \lambda=1..6}$).

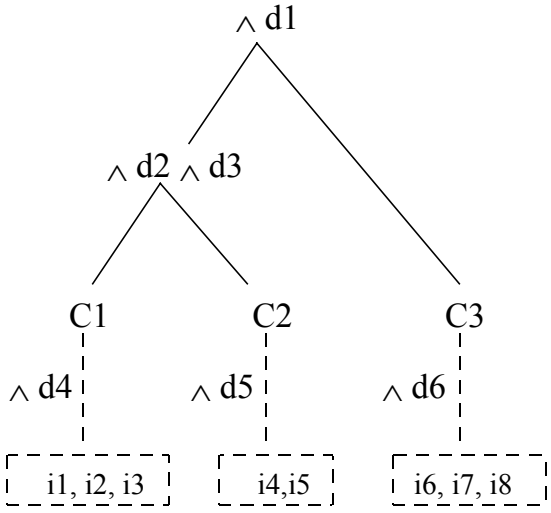


Fig.VI.2.3. Représentation de la classification.

VI.2.4- Connaissances classées

Chaque discipline scientifique recourt à des classifications pour découvrir les principaux regroupements. le comportement vis-à-vis de certains paramètres retenus se caractérise suivant les domaines. En Botanique, il permet de mettre en évidence des sous-espèces d'une même variété. En Reconnaissance des formes, il permet de classer les différents types d'écriture, En Biologie, il catalyse la description et l'organisation des connaissances sur le

monde vivant. Enfin, en Recherche documentaire, il favorise la recherche (inductive) de connaissances utiles de préférence pas trop imprécises [KODRATOFF, 2000a-b] ; etc.

Pratiquement, les classifications scientifiques visent deux objectifs principaux : – assurer l'identification des items inconnus et la prédiction des attributs des items intégrés en fonction de leur place dans la classification, et – permettre l'organisation et la structuration systématique des connaissances dans un domaine.

Le choix des critères de différenciation est un aspect fondamental dans une procédure de classification, car les enjeux de la classification ressortent de quatre grandes problématiques :

- Choix du critère : critère unique ou critères multiples (hiérarchisé ou non), regroupement des individus par proximité (empiriquement ou constatations scientifiques),
- Choix de la méthode : méthode empirique et construite (en fonction des observations), ou intuitive et traditionnelle (constructions conventionnelles, ou naturelles, etc.),
- Détermination du projet : définition des ensembles naturels, des catégories minimales ou inventaire ordonné,
- Structure effective de la classification : structure hiérarchique rigoureuse et univoque ou non.

Le choix de la méthode de construction paraît lié au but poursuivi. En revanche, le choix de « bons » critères ou de « bonnes » structures de résultat n'ont pas de réponse satisfaisante.

Une classification ne peut pas être « bonne » ou « mauvaise » dans l'absolu, mais elle pourra être utile, adéquate, profitable, intelligible, nouvelle, modifiant le comportement d'un agent particulier en fonction de l'usage que l'on veut en faire.

VI.2.5- Principe de raisonnement par classification

Nous venons de montrer, d'un point de vue cognitif, que les classifications servent deux fonctions principales : – la compréhension et – la mémorisation.
D'un point de vue pratique, les classifications visent deux objectifs principaux :

- l'identification d'individus,
- la prédiction des caractéristiques et l'organisation des connaissances d'un domaine [FELDMAN, 96 ET 98a-b].

Il nous paraît utile de raffiner la notion d'organisation des connaissances. Selon le principe de raisonnement par classification on peut tenir compte du fait que l'on peut organiser des connaissances dans un but de découverte ou non.

Nous distinguons donc trois types d'utilisation des classifications (*Fig. VI.2.5.*) :

- La prédiction de valeurs de caractéristiques inconnues : production de règles classificatoires (si entité *e* possède les caractéristiques {*x,y,z,…*} alors *e* possède la caractéristique *w*).
- L'organisation des connaissances : organisation indépendamment de tout pouvoir prédictif. On s'intéresse à la classification (ses résultats) et aux entités qui ont permis cette construction. En recherche documentaire, on peut distinguer deux utilisations, à savoir la recherche d'information (connaissances) et la navigation dans les hypermédias.

– L'aide à la découverte scientifique : la classification devient un objet d'étude. La comparaison entre les différentes classifications pourra faire l'objet des découvertes.

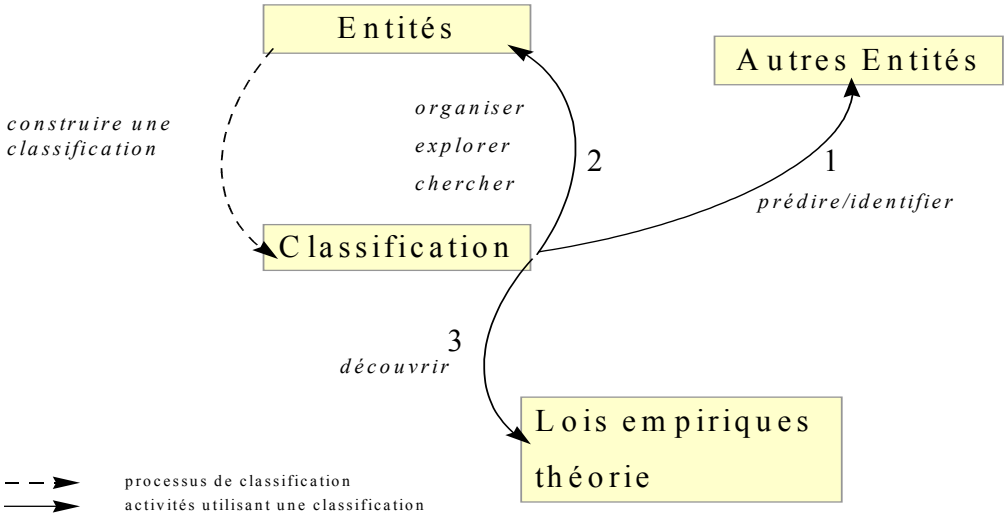


Fig. VI.2.5. Les utilisations de classifications : prédire, organiser et découvrir [BOURNAUD, 96] p.32.

VI.2.6- Consensus entre logique et techniques : la connaissance

L'extraction de connaissances à partir de données symboliques ou à partir de textes introduisent des notions nouvelles [DIXON, 97] : derrière les méthodes d'analyse devrait se trouver une interprétation qui modifie la nature des calculs effectués,

> « *Le fait que la connaissance soit nouvelle n'exclut pas qu'elle contredise la connaissance existante.* » [KODRATOFF, 2000a] p.12.

Dans le contexte de l'extraction des connaissances textuelles (ECT), certains auteurs la définissent comme étant de la recherche documentaire, soit de l'extraction d'information. D'autres, comme étant de la recherche inductive de connaissances utiles (pour des buts précis) dans les corpus textuels.

Du fait que les mesures, dans le cadre de nos travaux, dépendent des hypothèses reliées à la logique, nous considérons par ces faits que l'extraction des connaissances à partir de textes exige une meilleure coordination entre la logique et les techniques d'extraction (classification automatique). Mesure qui semble pertinente et qui va permettre de relier la sémantique des règles appliquées à la sémantique des associations textuelles intrinsèquement liées à la logique.
Cette hypothèse va permettre d'élever le niveau d'abstraction symbolique dans un système TALN. Celui-ci serait capable d'analyser « des discours » afin de retrouver des thèmes. Ces derniers forment des connaissances qui se retrouvent explicitement dans le texte analysé, et qui seraient ouvertes à des thèmes non explicites selon des mesures de corrélation et de causalité.

VI.3- Autonomie logique du syntagme nominal

En règle générale, dans le fonctionnement d'un système d'information documentaire ou non (bases de données textuelles ou bibliographiques, centre de documentation, etc.), la description de contenu d'un document ne se réduit, ni à un indice de classification (indices de

similarités vis à vis d'un tel ou tel critère), ni à une liste de mots (mots-clés), ni à une liste de termes (unités syntagmatiques dont la taille est supérieure à un mot).

La notice bibliographique qui accompagne en principe chaque document comporte d'autres informations plus complètes par rapport au titre du document, son auteur, sa production, etc. Les résumés réalisés sur le document coïncident sur ce type d'information complète. Ces éléments sont très porteurs de références à la réalité extra-linguistique et permettent d'offrir de la pertinence sur les informations.

Le processus de recherche d'informations sera évidemment centré sur la mise en oeuvre du mécanisme de référence à la réalité extra-linguistique [LAINÉ, 82], [DE BRITO, 91], [KURAMOTO, 95-99]. Une analyse identique à celle du document (texte intégral) ou de la notice bibliographique attachée au document (champs textuels de type résumé) permet d'identifier dans la question exprimée en langage naturel un ou plusieurs syntagmes nominaux [LALLICH, 86] [JOUVE, 99] [JONES, 91] [LANCASTER, 91-93].

VI.3.1- Relations entre constituants d'un syntagme nominal

Au sein du groupe SYDO, le descripteur doit être un signe doté de valeur référentielle. Les syntagmes nominaux ne sont pas des signes sans référence. A l'opposé des mots isolés, ils sont composés de mots dans un ordre donné et souvent avec des liaisons syntaxiques. Ces mots qui composent le syntagme nominal ne constituent plus des ensembles de prédicats libres ou de propriétés. Ainsi, dans le contexte du syntagme nominal, chaque mot a son rôle bien défini, avec une signification spécifique, et par conséquent, le syntagme nominal se définit comme étant la plus petite unité du discours porteuse d'une valeur référentielle. Dans le contexte de connaissance de type « information », le syntagme nominal est la plus petite unité d'information dans un texte.

Cette opposition entre les mots en tant qu'unité du lexique hors discours, et les mots qui font partie du discours, est bien caractérisée de manière plus spécifique pour le syntagme nominal. D'abord le mot, en tant que mot de la langue, qu'unité du lexique, est au niveau N. Avant qu'il fasse partie du syntagme nominal le mot passe par un niveau intermédiaire (N') où il prend ses valeurs sur l'univers du discours. La distinction entre ces deux niveaux est qu'au niveau N, le mot n'a qu'un ensemble de propriétés. Il ne désigne aucun objet quel qu'il soit. Il n'y a donc aucune référence à un objet du monde réel. Tandis que lorsqu'il est au niveau N', il désigne un objet ou au moins il fait référence à une classe d'objets.

Le syntagme nominal est la mise en oeuvre de deux organisations logiques différentes. Au niveau N, on a l'intervention de la logique intensionnelle. Cette logique est « *sans référentiel et sans classe, constituée de relations et de propriétés envisagées indépendamment de quelque objet que ce soit* » [LE GUERN, 91]. Tandis qu'au niveau N'' et en partie au niveau N', ils relèvent de la logique extensionnelle, car le prédicat libre prend ses valeurs sur un univers du discours. Là on peut envisager une classe d'objets.

« *Le passage du niveau N au niveau N' correspond à la prise en compte d'un univers donné, au surgissement de la référence, à la possibilité de déterminer des classes, au moins virtuelles ; c'est le basculement de la logique intensionnelle à la logique extensionnelle ; c'est la mise en relation des mots et des choses.*» [LE GUERN, 91].

Le niveau N'' correspond à la fermeture logique du syntagme nominal, caractérisée par l'ajout d'élément déterminant, c'est la complète référence à un objet donné.

VI.3.2- Organisations naturelles de syntagmes nominaux

Les syntagmes nominaux ont une organisation naturelle. Dans un sens, ils ont un rapport d'*emboîtement* les uns avec les autres, ce qui permet de les classifier en des niveaux informationnels distincts. Et dans l'autre, ils ont un rapport de *ramification*, dans le cas où le syntagme nominal est avec une double rection. Cette dernière propriété permet d'ordonner et de distinguer les classes d'informations : *structure d'arbre des classes d'informations*. Ces caractéristiques permettent de construire une architecture de connaissances et d'exploiter les données (= les SN) au moyen de la navigation dans des structures arborescentes. Par la superposition des données sur les SN avec les centres de syntagmes (= les N), la navigation dans les structures s'intègre dans une architecture treillis de données.

Pour montrer la caractéristique d'emboîtement, on présentera un exemple, d'un syntagme nominal de troisième niveau. On utilise le mot *niveau* pour indiquer l'ordre d'extraction des syntagmes nominaux. En effet, la grandeur du niveau est inversement proportionnelle à l'ordre d'extraction.

Exemple : Le projet de loi de la politique économique de la France. (*Figure VI.3.2a*)

[Le projet de loi de [la politique économique de [la France] ^{SN1}] ^{SN2}] ^{SN3}

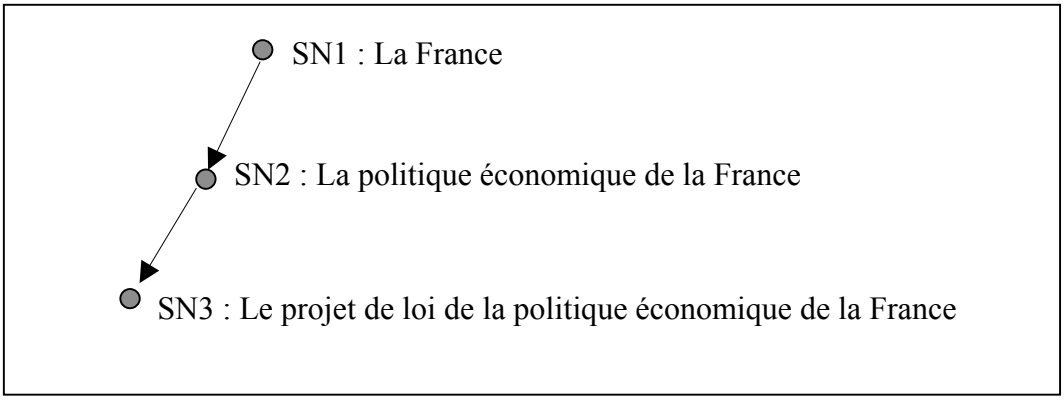


Figure VI.3.2a. Exemple d'emboîtement de syntagmes nominaux

Le rassemblement de tous les syntagmes nominaux dans une base de données permettra de construire une structure arborescente.

Pour montrer la caractéristique de ramification, on présentera un exemple, d'un syntagme nominal avec une double rection. Nous présentons quelques exemples pour une meilleure visualisation de cette proposition.

Exemple 1 :

« l'acceptation de l'information stratégique dans la définition de l'avenir de l'entreprise » [KURAMOTO, 99] (*Fig. VI.3.2b.*).

[l'acceptation de [l'information stratégique]^{SN1} dans [la définition de [l'avenir de [l'entreprise]^{SN1}]^{SN2}]^{SN3}]^{SNmax}.

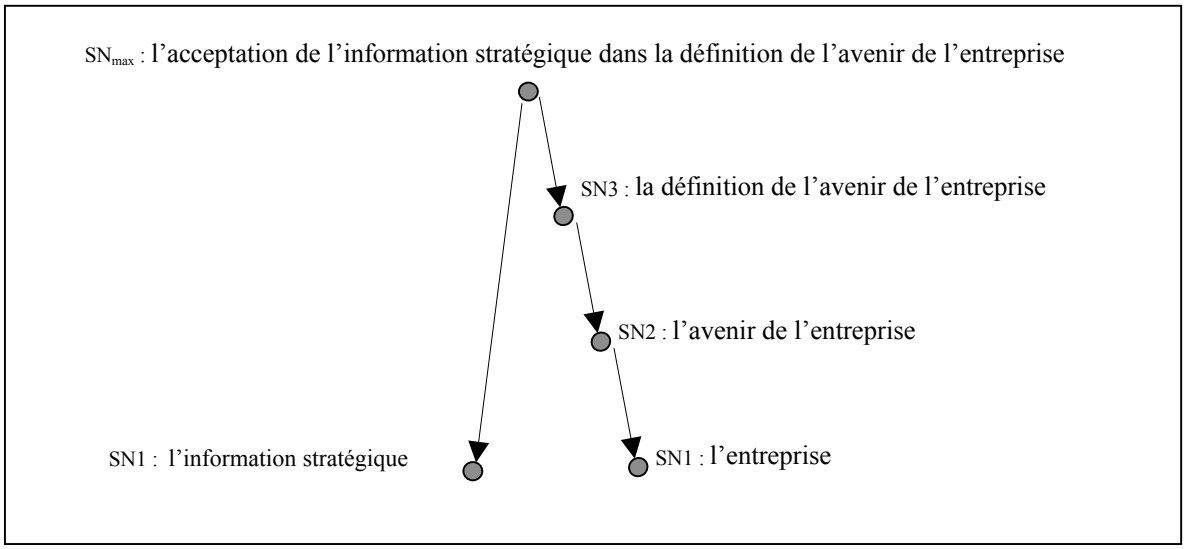


Fig. VI.3.2b. Exemple 1 d'arborescence de syntagmes nominaux

Etant donné que ce syntagme contient une double rection, il faut repérer le niveau du syntagme nominal maximal par rapport au syntagme nominal plus petit de chaque rection. C'est-à-dire, le cas d'appartenance à deux niveaux distincts à cause de ses rections. Ainsi, nous avons le calcul suivant :

- (l'information stratégique) : *Niveau 1*
- (l'entreprise) : *Niveau 1*
- (l'avenir de (l'entreprise)) : *Niveau 2*
- (la définition de (l'avenir de (l'entreprise))) : *Niveau 3*
- (l'acceptation de (l'information stratégique) dans (la définition de (l'avenir de (l'entreprise))) : *Niveau 0 =Niveau maximal*

Exemple 2 :

« Les rapprochements faits avec l'apparition de certaines comètes par les scientifiques. », (résumé INA , *réf.* DL T 19950108 5E 002), (*Fig.VI.3.2c.*).

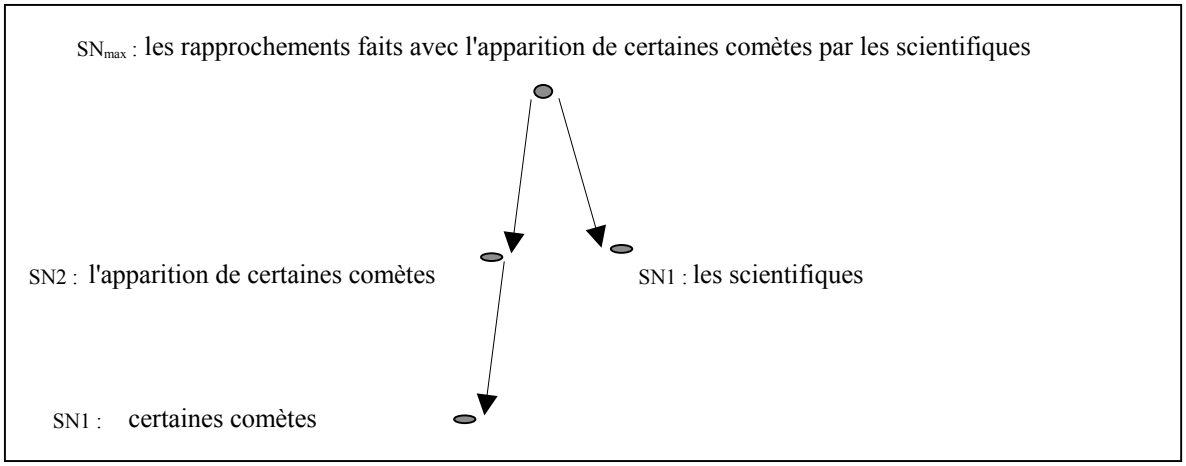


Fig. VI.3.2c. Exemple 2 d'arborescence de syntagmes nominaux.

VI.3.3- Schémas de sélection et d'interrogation

➢ Schéma de sélection

La différenciation des prédicats intensionnels (ou libres) aux prédicats extensionnels (ou saturés, les SN), permet de résoudre le problème majeur lié à l'extraction de l'information. La distinction des éléments, qui ont des propriétés prédicatives intensionnelles aux éléments qui ont des fonctions référentielles comme les syntagmes nominaux, permet de fournir une approche nouvelle de type référentielle (ou logique extensionnelle) dans le schéma de construction d'un système d'information (*Fig.VI.3.3a.*).

Dans notre application, la base comportera des connaissances et des faits stockés :

- Prédicats intensionnels : nous avons vu qu'un SN peut se définir comme une suite de prédicats libres construits autour d'un nom. Ce nom fait directement référence à un élément extra-linguistique. Le nom employé comme centre du syntagme fera le lien à sa référence lors de son instanciation (saturation).
- Prédicats extensionnels non saturés : Le SN est presque toujours le thème en faisant référence à la correspondance entre les SN extraits d'un texte par l'analyseur et les descripteurs issus d'une indexation manuelle. Les SN non saturés correspondent aux SN englobés/emboîtés dans d'autres SN de niveau supérieur.
- Prédicats extensionnels saturés : Ils correspondent aux SN qui contiennent tous les autres SN de niveau inférieur. Un SN de ce type représente le thème générique et complet dans le texte.

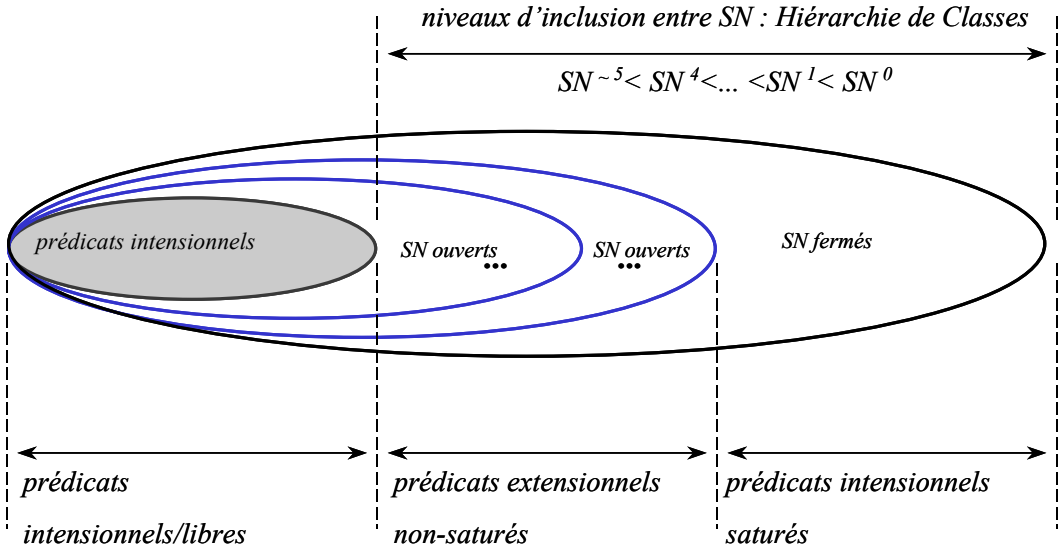


Fig.VI.3.3a. Approche logique Intensionnelle/Extentionnelle dans la construction d'un système d'information

Aussi bien des relations entre ces connaissances :

♦ Relation d'inclusion

Cette relation d'appartenance qui détermine si le SN est inclus dans d'autres SN permet de le situer dans la hiérarchie des classes SN (*Fig. VI.3.3b.*). Une hiérarchie facile à gérer pour une structure en arbre.

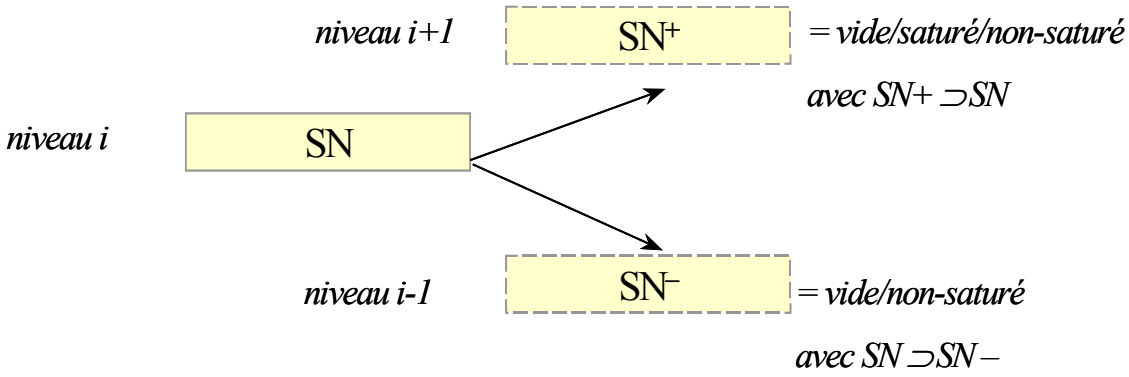


Fig. VI.3.3b. Relations d'appartenance dans un SN

♦ Relation de générique-spécifique

Le prédicat libre (N) dans le syntagme nominal est un élément appartenant à la logique intensionnelle. Ce prédicat N ne peut construire un objet de discours, mais comme trait d'une classe pour accéder à ses éléments ou SN (*Fig. VI.3.3c.*). Ce prédicat est souvent représenté par un <*nom*> comme centre du syntagme nominal et contribuant à la description d'une classe d'objets (ou point d'accès).

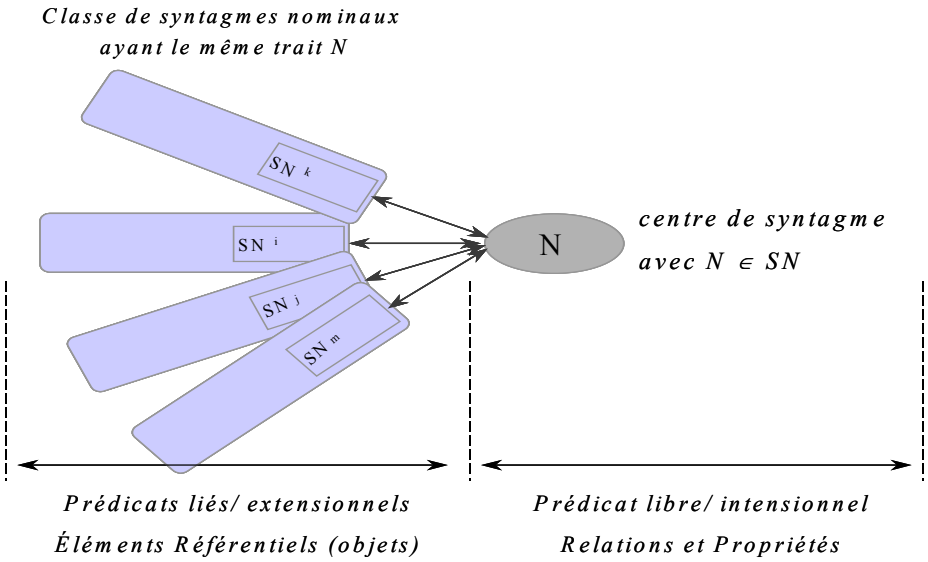


Fig. VI.3.3c. Centre de syntagme (prédicat intensionnel) comme point d'accès aux SN

➢ Schéma d'interrogation

Le travail réalisé concerne l'identification des parties du discours construites autour du nom. Ces parties sont porteuses de référence aux objets de l'univers du discours. Elles sont celles qu'il faut identifier aussi bien dans l'opération d'indexation des documents que dans l'opération d'indexation de la requête de l'usager.

Dans ce contexte, le schéma d'interrogation de la base informationnelle (connaissances) consiste à retrouver les syntagmes nominaux de la requête qui sont présents dans la base. Bien entendu, les documents qui répondent le mieux à la requête sont ceux identifiés par des SN saturés, bien moins que par ceux identifiés par les SN non-saturés, et moins par ceux identifiés par les prédicats intensionnels ou les nominaux (*Fig. VI.3.3d.*).

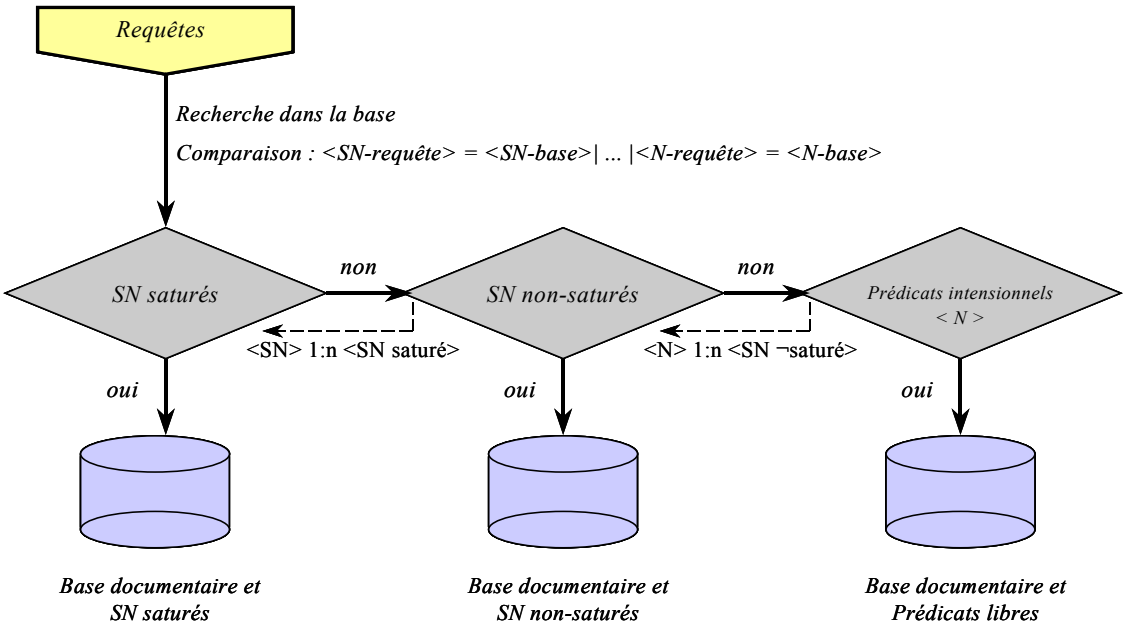


Fig. VI.3.3d. Schéma d'interrogation de la Base documentaire

VI.4- Représentation profonde

L'information enregistrée dans la base de donnée n'étant pas de même nature que celle qui caractérise habituellement les systèmes documentaires. Ainsi, la stratégie de recherche d'informations sera totalement différente, car les informations enregistrées concernant les documents et les requêtes (*Fig. VI.4*) sont analysées et structurées par les mêmes modules de traitements (Modules de l'Analyseur Morpho-syntaxique).

Les informations finales à manipuler et à comparer entre elles ont l'avantage d'être structurées d'une manière plus fine et plus riche que celles dont on dispose dans les systèmes classiques.

Une information « *plus fine* » est la conséquence, d'une part, de la relation d'inclusion entre les syntagmes nominaux et, d'autre part, de la relation générique-spécifique entre le centre et son syntagme nominal. « *Plus riche* », elle est la conséquence du modèle linguistique (Modèle SYDO) adopté pour l'indexation documentaire basé sur l'extraction des syntagmes nominaux. Une description détaillée des structures morpho-syntaxiques du discours étant une bonne représentation de cette sémantique.

L'ensemble des associations reliant les références (SN) avec les traits (N) des classes d'objets, *c'est-à-dire* relations SN–SN et relations N–SN, décrit correctement l'organisation morpho-syntaxique des différentes parties du discours et permet d'établir un réseau cohérent d'informations structurées dans la base.

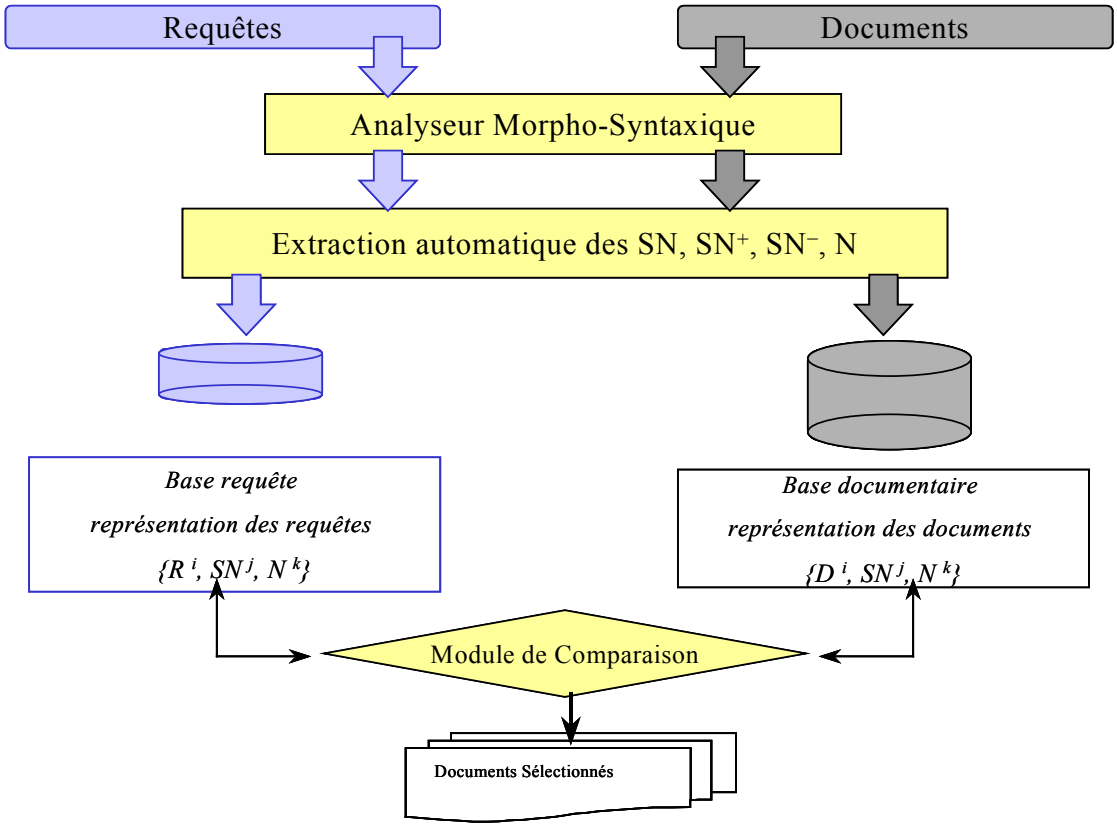


Fig. VI.4. Modèle du SRI basé sur la manipulation des SN

- **Exemple :** Fichier inverse (Indexation automatique de la Base des notices)

Nous présentons dans le tableau suivant (*Tab.VI.4.*) un échantillon du fichier inverse de la base des notices documentaires. Les informations constituées permettent d'identifier les unités du discours de manière logique (les SN et leurs emboîtements, les N) en relation avec les unités documentaires (les notices DL_notice) avec le coefficient d'analyse du SN dans la phrase (coeff).

coeff	DL_notice	N	SN	SN+	SN–
...					
83	137337.001	atelier	le atelier de construction aéronautique de Matra	_	Matra
100	262714.022	bande	la bande dessinée	_	_
92	87820.001	but	Le but de cette expédition	_	cette expédition
94	842.001	chercheurs	des chercheurs dans le nouveau Québec	une équipe de des chercheurs dans le nouveau Québec	le nouveau Québec
100	262714.022	cinéma	le cinéma	_	_
91	57399.001	compagnie	la compagnie de les dockers	le fonctionnement autogestionnaire de la compagnie de les dockers	les dockers
100	422972.001	conception	sa conception secrète sous l' occupation à son succès commercial	_	l' occupation à son succès commercial

coeff	DL_notice	N	SN	SN+	SN–
100	262714.022	cybernétique	la cybernétique	_	_
100	262714.022	développement	le développement de la intelligence artificielle à les robots industriels	_	la intelligence artificielle à les robots industriels
91	57399.001	dockers	les dockers	la compagnie de les dockers	_
94	842.001	documentaire	Ce documentaire	_	_
100	378275.001	documentaire	un documentaire en deux parties	Ce premier volet de un documentaire en deux parties	_
100	58230.001	domaines	les domaines de les réseaux informatiques	_	les réseaux informatiques
83	137337.001	enquête	la enquête	cette première partie de la enquête	_
94	842.001	équipe	une équipe de des chercheurs dans le nouveau Québec	les travaux menés par une équipe de des chercheurs dans le nouveau Québec	des chercheurs dans le nouveau Québec
94	378275.001	espace	l' espace	le terrien face à l' espace	_
92	87820.001	expédition	cette expédition	Le but de cette expédition	_
94	378275.001	face	le terrien face à l' espace	_	l' espace
91	57399.001	fonctionnement	le fonctionnement autogestionnaire de la compagnie de les dockers	_	la compagnie de les dockers
87	422972.001	histoire	la histoire de le véhicule	_	le véhicule
90	378275.001	histoire	la histoire de les pionniers de l' espace	_	les pionniers de l' espace
100	262714.022	intelligence	la intelligence artificielle à les robots industriels	le développement de la intelligence artificielle à les robots industriels	les robots industriels
92	58230.001	laboratoires	leurs laboratoires	les locaux de leurs laboratoires	_
83	262714.022	littérature	la littérature	_	_
92	58230.001	locaux	les locaux de leurs laboratoires	_	leurs laboratoires
92	87820.001	matière	la matière cosmique	_	_
91	57399.001	métier	leur métier	_	_
100	422972.001	occupation	l' occupation à son succès commercial	sa conception secrète sous l' occupation à son succès commercial	son succès commercial
90	378275.001	pionniers	les pionniers de l' espace	la histoire de les pionniers de l' espace	l' espace
100	7359.001	professeur	Le professeur Michel IMBERT	_	_
100	58230.001	réseaux	les réseaux informatiques	les domaines de les réseaux informatiques	_
90	262714.022	robots	les robots	tout le univers de les	_

coeff	DL_notice	N	SN	SN+	SN–
				robots	
100	262714.022	robots	les robots industriels	la intelligence artificielle à les robots industriels	_
92	58230.001	site	le site en extérieur	_	_
100	422972.001	succès	son succès commercial	l' occupation à son succès commercial	_
94	842.001	travaux	les travaux menés par une équipe de des chercheurs dans le nouveau Québec	_	une équipe de des chercheurs dans le nouveau Québec
90	262714.022	univers	tout le univers de les robots	_	les robots
87	422972.001	véhicule	le véhicule	la histoire de le véhicule	_
100	7359.001	yeux	nos deux yeux	_	_

Tab.VI.4. Echantillon du fichier inverse de la base de notices.

VI.5- Choix et limitations du modèle représenté

Pour un document donné, l'ultime élément informatif au sens où il renvoie à des éléments référentiels sont de type SN. Cependant, il est parfois difficile de faire coïncider les mêmes types d'objet dans l'univers de l'utilisateur (ses requêtes) lors de l'interrogation. De même que ceux existant dans la base. Les rapprochements actuels des objets utilisateur avec la base se font selon les niveaux logiques extensionnel puis intensionnel.

Le processus de recherche (*Fig.VI.3.3d. et Fig.VI.4.*) consiste à rechercher les objets ayant des caractéristiques communes entre requêtes et base, *c'est-à-dire* appartenant à une même classe ou sous-classe d'objets (SN saturés et non saturés). Dans le cas échéant, ce processus consiste à retrouver des points d'accès communs aux classes d'objets, c'est-à-dire des prédicats intensionnels communs.

Il est évident qu'un locuteur qui a des idées précises sur le sujet, a une description qui ne se présente pas toujours sous la même forme. A l'exception des objets uniques et universels qui appartiennent à tous les univers de discours. Cela est vrai en pratique pour les noms propres et de nombreux objets spécifiques. Il reste l'autre catégorie des objets non spécifiques.
Comment remédier à ce problème ?

La manipulation progressive des objets conduit à l'analyse des figures et des formes, tandis que la notion de l'objet demeure inchangée. Dès lors, le langage apparaît comme un prolongement de l'objet dans un espace qui lui accorde cette multiplicité (plusieurs description d'un même objet). De l'avis de Saussure, « *dans la langue il n'y a que des différences.* » ; Dans ce cadre, l'unité linguistique n'est pas un signe, mais une valeur, et cette valeur est le résultat de relations complexes intervenant à plusieurs niveaux que celui d'une correspondance simpliste entre signifiant et signifié à l'intérieur du mot et du morphème.
Lorsque le signifiant et le signifié sont « pris séparément », seule leur combinaison « fondatrice du signe » est un fait positif : avec les règles de réécriture du modèle syntagmatique – analyse en constituants immédiats et règles engendrant l'indicateur syntagmatique de la structure profonde.

L'approche proposée (*Fig.VI.5.*), pour résoudre le problème d'échec de la recherche d'information dans la base, consiste à expérimenter des recherches basées sur les synonymes des prédicats intensionnels existant dans les requêtes.
Les synonymes Ns, qui sont des prédicats libres synonymes à ceux de la requête, permettront de retrouver dans la base les syntagmes ayant pour centre Ns et par conséquent les références aux documents de la base.

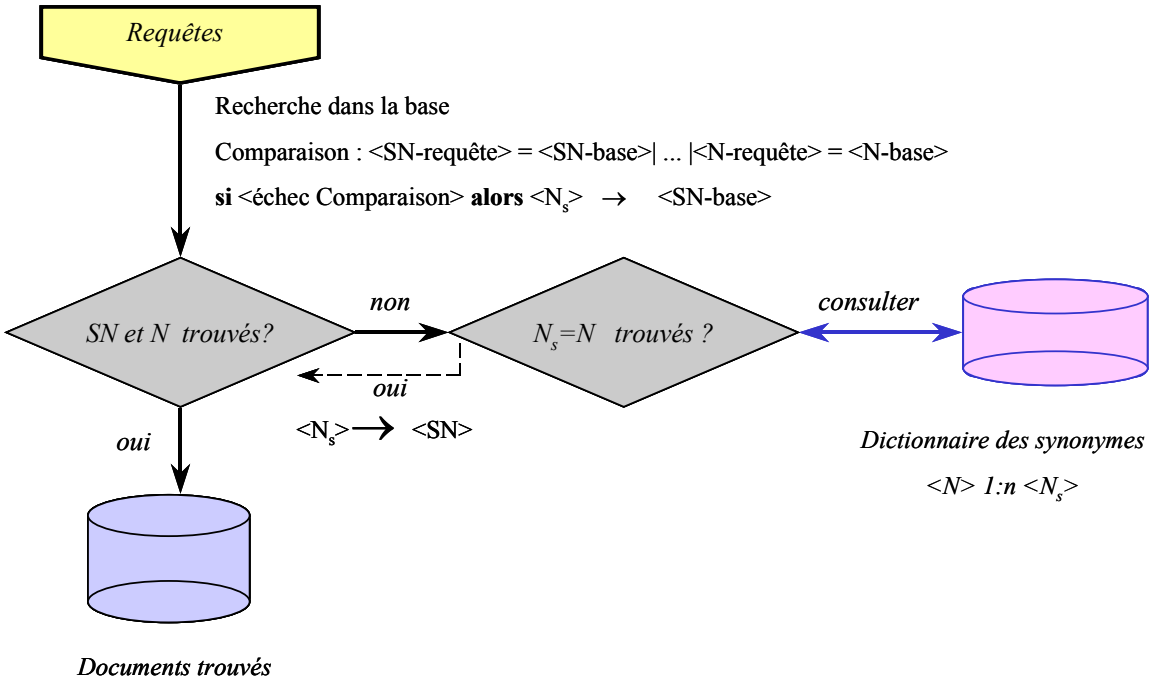


Fig. VI.5. Nouveau Schéma d'interrogation de la Base documentaire : recherche des synonymes.

Exemple : indexation automatique d'une requête

Le texte de la requête peut se présenter (ou formuler) comme un texte résumé dans une notice INA. Celui-ci a pour thème « Airbus industrie » (*réf.* Notice INA : DL T 19950107 M6 007). Le processus d'indexation de la requête (*Tab.VI.5.*) fait appel à notre analyseur morpho-syntaxique.

Titre	Airbus Industrie
REQ001	Pour les 20 ans d'AIRBUS INDUSTRIE, Frédéric BORSU retrace l'historique du consortium européen au moyen de documents d'archives. Du A 300B au A340, toutes les phases nous sont montrées pour comprendre les difficultés techniques et économiques qu'a pu rencontrer AIRBUS INDUSTRIE.
REQ002	AIRBUS INDUSTRIE parallèlement à l'Europe qui est en pleine mutation (chute du mur de Berlin, Tunnel sous la Manche...) prépare son avenir avec le A 321. Le A 300 B fit son vol inaugural, le 28 octobre 1972 avec Bernard ZIEGLER comme pilote d'essai. Mais toute la période de construction fut semé d'embauches tant économiques, politiques que techniques. La construction de la voilure fut construite à Chester en Angleterre, la section principale du fuselage à Hambourg en Allemagne, le cockpit et la section centrale à Saint-Nazaire et à Toulon par l'Aérospatial. Le plus dur fut d'acheminer les différentes parties à BLagnac pour l'assemblage du prototype. Les voies aériennes et routières fut empruntées non sans mal. Les tests draconiens du programme de certification sont passés avec réussite. Le succès commercial de l'Airbus A300B a permis au consortium européen d'élargir sa gamme d'appareils. Suivant les mêmes techniques de recherche, de construction et d'assemblage, AIRBUS INDUSTRIE construira les modèles A310, A320, A330 et A340.

coeff	ref_req	N	SN	SN+	SN–
…					
87	REQ002	A 321	le A 321	son avenir avec le A 321	_
33	REQ002	A 300	Le A 300	_	_
33	REQ002	Airbus	le Airbus A300B	Le succès commercial de le Airbus A300B	_
33	REQ002	Airbus	le Airbus A300B	Le succès commercial de le Airbus A300B	_
75	REQ001	ans	les 20 ans de AIRBUS	_	_
87	REQ002	avenir	son avenir avec le A 321	_	le A 321
44	REQ002	B	300 B	_	_
100	REQ002	cockpit	le cockpit	_	_
94	REQ001	consortium	le consortium européen à le moyen de des documents de archives	le historique de le consortium européen à le moyen de des documents de archives	le moyen de des documents de archives
83	REQ002	construction	La construction de la voilure	_	la voilure
94	REQ001	historique	le historique de le consortium européen à le moyen de des documents de archives	_	le consortium européen à le moyen de des documents de archives
94	REQ001	moyen	le moyen de des documents de archives	le consortium européen à le moyen de des documents de archives	des documents de archives
41	REQ002	octobre	le 28 octobre 1972	_	_
42	REQ002	période	la période de construction	_	_
92	REQ002	programme	le programme de certification	Les tests draconiens de le programme de certification	_
100	REQ002	section	la section centrale à Saint-Nazaire	_	Saint-Nazaire
33	REQ002	succès	Le succès commercial de le Airbus A300B	_	le Airbus A300B
83	REQ002	voilure	la voilure	La construction de la voilure	_
…					

Tab.VI.5. Echantillon du « fichier inverse » d'indexation du texte de la requête.

VI.6- Evaluations

Autour de la question de l'indexation automatique, nous avons choisi de donner à ce travail un caractère pluridisciplinaire. Cet aspect a rendu possible la description théorique et pratique d'un modèle global d'analyse syntaxique du langage (français) à un certain type d'applications : les fondements théoriques de l'indexation automatique employant une approche linguistique.

VI.6.1- Evaluation logique

Le processus d'indexation consiste en une extraction d'unités de discours. Antérieurement, nous avons spécifié le rôle de ces unités extraites des textes analysés en nous appuyant sur le modèle des chaînes de référence.

D'un point de vue logique, les unités d'indexation (les *syntagmes nominaux* comme descripteurs) extraites sont pourvues des propriétés suivantes :

– Unités qui sont dotées d'un pouvoir référentiel leur accordant le statut d'identifier un ou plusieurs objets du domaine (dans l'univers du discours)

– Unités qui présentent la propriété de dénomination d'objet et de classe : description d'une relation directe vers un objet de la classe, ou indirectement, d'une relation vers une classe d'objets ;

– Unités qui ont le pouvoir d'une description quantifiée : La nature du descripteur engage de référer des unités de discours, non des unités de langue. Une opposition que Benveniste établissait dans « Sémiologie de la langue », qui serait pertinent pour le lexique, et le « mode sémantique », dont relèveraient les descripteurs [LE GUERN, 91].

- **Exemple :** Les unités du discours

Nous présentons dans le tableau suivant (*Tab.VI.6.1.*) un échantillon de connaissances qui sont extraites du fichier inverse de la base des notices indexées par leurs résumés. Ces connaissances sont composées des unités du discours qui permettent de reconstituer les informations textuelles de manière logique, les SN avec leurs emboîtements et les N (centre nominal d'un SN).

N	SN	SN+	SN–
…			
aventure	la aventure vécue par Jean-Louis ETIENNE	_	Jean-Louis ETIENNE
conception	sa conception	_	_
conception	sa conception secrète sous l' occupation à son succès commercial	_	l' occupation à son succès commercial
développement	le développement de la intelligence artificielle à les robots industriels	_	la intelligence artificielle à les robots industriels
documentaire	un documentaire en deux parties	Ce premier volet de un documentaire en deux parties	deux parties
domaines	les domaines de les réseaux informatiques	_	les réseaux informatiques
équipe	une équipe de des chercheurs dans le nouveau Québec	les travaux menés par une équipe de des chercheurs dans le nouveau Québec	des chercheurs dans le nouveau Québec
espace	l' espace	le terrien face à l' espace	_
intelligence	la intelligence artificielle à les robots industriels	le développement de la intelligence artificielle à les	les robots industriels

N	SN	SN+	SN–
		robots industriels	
occupation	l' occupation à son succès commercial	sa conception secrète sous l' occupation à son succès commercial	son succès commercial
professeur	Le professeur Michel IMBERT	_	_
réseaux	les réseaux informatiques	les domaines de les réseaux informatiques	_
robots	les robots industriels	la intelligence artificielle à les robots industriels	_
succès	son succès commercial	l' occupation à son succès commercial	_
travaux	les travaux menés par une équipe de des chercheurs dans le nouveau Québec	_	une équipe de des chercheurs dans le nouveau Québec
volet	Ce premier volet de un documentaire en deux parties	_	un documentaire en deux parties
yeux	nos deux yeux	_	_
yeux	Derrière nos deux yeux	_	_
…			

Tab.VI.6.1. Echantillon de connaissances : des unités du discours.

VI.6.2- Evaluation linguistique

Une représentation linguistique du descripteur comme unité de discours consiste à l'identifier explicitement comme un syntagme nominal. La notion de SN sous-tend explicitement un modèle syntaxique (grammaire de réécriture du SN), et implicitement un comportement interprétatif du SN en discours (modèle logico-sémantique) : relation référentielle autonome.

Dans la représentation linguistique proposée, la syntaxe est comprise comme un mode d'organisation de l'interprétation :

– la morphologie du descripteur, qui est définie formellement selon le point de vue de l'indexeur et celui de l'utilisateur,

– la syntaxe du descripteur, qui se manifeste par ses dispositions à l'implémentation automatisée et à l'extraction automatique des unités de discours.

Pour illustrer cet aspect sur l'évaluation linguistique, nous procédons dans ce qui suit par une étude comparative de trois méthodes d'indexation :

– indexation manuelle : (méthode intellectuelle) extraction manuelle des SN,

– indexation automatique : emploi du noyau d'indexation pour l'extraction des SN,

– indexation INA : emploi d'un thésaurus spécialisé de l'audiovisuel.

➢ Etude comparative des méthodes d'indexation

Cette étude concerne l'analyse d'un corpus de cent (100) cotices documentaires de l'INA. Chacune comporte un résumé court (chapeau), un résumé détaillé (résumé) et les descripteurs INA (DES). L'objet de cette étude consiste à observer trois méthodes d'indexation : manuelle, automatique et par thésaurus. L'étude appliquée au même corpus consiste à confronter les trois méthodes d'indexation devant le paramètre descripteur qu'est le syntagme nominal.

A. Indexation Manuelle du corpus

chapeau :

par 1 notice :	nbr PHR	nbr SN	SN/PHR
	2	14	7

ce qui correspond (en moyenne) par 1 phrase à 7 SN

résumé :

par 1 notice :	nbr PHR	nbr SN	SN/PHR
	5.8	29.4	5

ce qui correspond (en moyenne) par 1 phrase à 5 SN

Ces résultats sont l'aboutissement d'une analyse manuelle (indexation intellectuelle) sur le corpus en extrayant les SN dans les résumés selon le modèle syntaxique des syntagmes nominaux.

B. Indexation Automatique du Corpus

chapeau :

par 1 notice :	nbr PHR	nbr SN	SN/PHR
	2	18.3	9.15

ce qui correspond (en moyenne) par 1 phrase à 9 SN

résumé :

par 1 notice :	nbr PHR	nbr SN	SN/PHR
	5.8	35.4	6.1

ce qui correspond (en moyenne) par 1 phrase à 6 SN

Ces résultats sont l'aboutissement d'une analyse automatique (noyau d'indexation) sur le corpus en extrayant les SN dans les résumés selon la grammaire de réécriture de notre modèle « cognitif » de la phrase incorporant la grammaire de réécriture du SN.

C. Indexation INA par thésaurus du corpus

chapeau et résumé :

par 1 notice :	nbr PHR	nbr DES	DES/PHR
	7.8	13	1.66

ce qui correspond (en moyenne) par 1 phrase à 2 DES.

Ces résultats sont l'aboutissement d'une analyse documentaire sur le contenu des documents primaires du corpus en indexant chaque contenu au moyen d'un thésaurus spécialisé de l'audiovisuel. Les documentalistes de l'INA ont la charge d'employer cette méthode y compris celle de la rédaction des résumés dans la constitution d'une notice.
La nature syntaxique des descripteurs (DES) est compatible au SN simple sans le déterminant (ou D-zéro).

D. Représentation comparative entre indexation Manuelle et Automatique

chapeau :

	nbr PHR	nbr SN	SN/PHR
Automatique	2	18.3	9.15
Manuelle	2	14	7

résumé :

	nbr PHR	nbr SN	SN/PHR
Automatique	5.8	35.4	6.1
Manuelle	5.8	29.4	5

L'élément commun entre ces deux méthodes d'indexation est le SN avec sa grammaire de réécriture. Les deux schémas comparatifs (*Fig.VI.6.2D*) permettent de montrer le paramètre de bruit qu'engendre la méthode d'indexation automatique. Dans ce cas d'analyse, l'indexation manuelle est la méthode de référence « idéale » pour faire la comparaison.

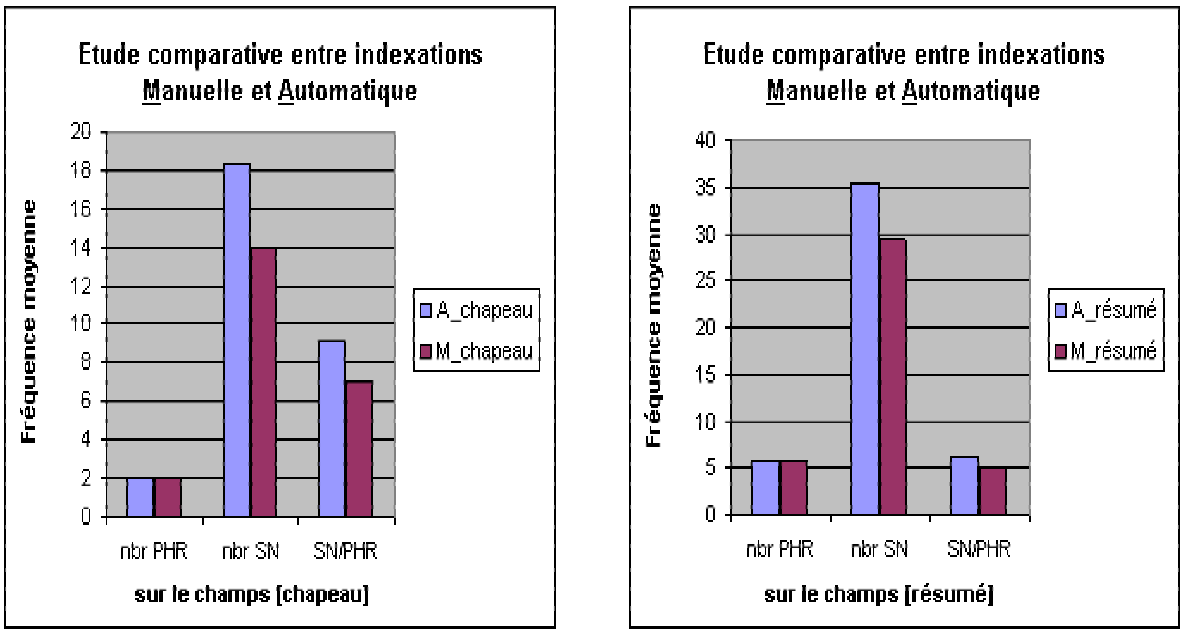


Fig.VI.6.2D. Graphes sur l'étude comparative entre indexation manuelle et automatique.

E. Représentation comparative entre les trois méthodes d'indexation

chapeau et résumé (par 1 notice):

	nbr PHR	nbr DES	DES/PHR
INA	7.8	13	1.66
MANUELLE	7.8	43.4	12
AUTOMATIQUE	7.8	53.7	15.25

Après l'étude de chaque méthode d'indexation (INA, manuelle, automatique), nous avons établis les résultats d'analyse (statistiques) propre à chacune sur le même corpus (100 notices). Le paramètre SN est le vecteur comparatif de cette étude.

Nos résultats statistiques sur le SN ont permis d'observer les caractères de bruit et de silence entre les méthodes d'indexation et d'établir des liens de rapprochement. Le schéma suivant (*Fig.*VI.6.2*E*) illustre ces phénomènes.

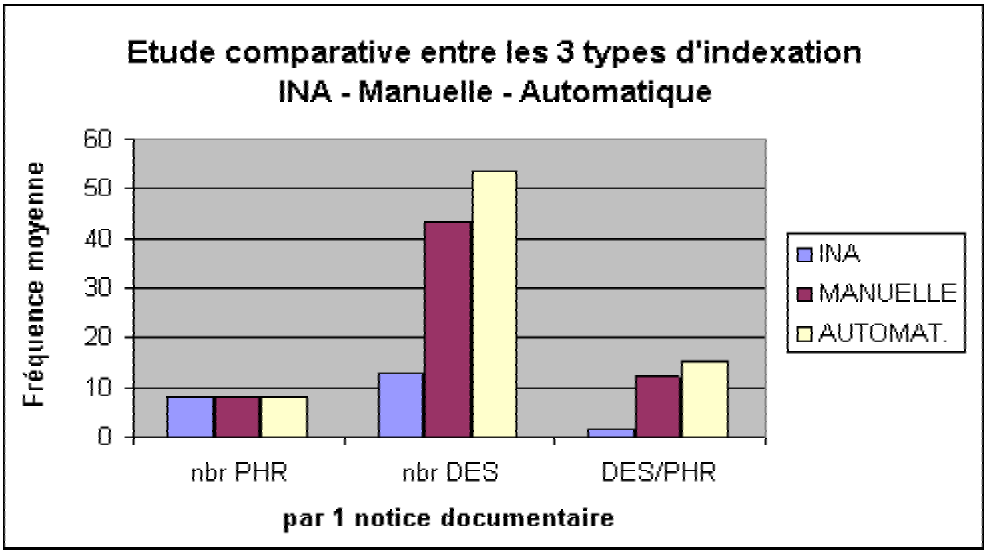


*Fig.*VI.6.2*E. Graphe sur l'étude comparative entre les 3 méthodes d' in*dexation.

F. Conclusion sur l'étude comparative

Trois méthodes d'indexation appliquées à 1 notice documentaire INA :

Indexation Automatique	nbr PHR	nbr SN	SN/PHR
chapeau	2	18.3	9.15
résumé	5.8	35.4	6.1
=	7.8	53.7	15.25

Indexation Manuelle	nbr PHR	nbr SN	SN/PHR
chapeau	2	14	7
résumé	5.8	29.4	5
=	7.8	43.4	12

Indexation INA par Thésaurus	nbr PHR	nbr DES	DES/PHR
INA	7.8	13	1.66

Cette étude comparative a démontré des aspects comparatifs concernant les trois méthodes d'indexation et leur degré de pertinence :

- l'indexation INA (par les descripteurs du thésaurus) comparée à l'indexation manuelle montre un caractère de *silence* que produit la méthode INA à celle manuelle,
- l'indexation manuelle (extraction des SN) comparée à l'indexation automatique montre un caractère de *bruit* que produit notre méthode automatique à celle manuelle.

Les explications qu'on peut apporter vis-à-vis de cette étude comparative sont les suivantes :

- notre approche pour l'indexation automatique produit des résultats très proches à l'indexation manuelle. Le bruit que produit l'indexation automatique est dû aux limites de l'analyseur morpho-syntaxique, c'est-à-dire analyse sans traitement de la coordination, sans prise en compte du phénomène de la rection verbale (attachement d'un SP) et les erreurs liées aux traitements de la ponctuation (ambiguïté, mauvaise ponctuation).
- L'approche INA basée sur l'indexation par thésaurus est réductrice devant les résultats obtenus manuellement ou automatiquement. L'INA réserve une recherche en « full text » par uni-termes sur le contenu de la notice pour remédier aux insuffisances des descripteurs.

VI.6.3- Evaluation classificatoire

La visée du projet est éminemment pratique une fois que les syntagmes nominaux sont extraits. La liste de tous les SN du corpus, accompagnés pour chaque élément de la liste des références de ses occurrences, de la liste de syntagmes liés (SN non saturés), et de son centre (N), est d'une utilité fondamentale à créer des classes de SN.

En effet, cette classification regroupe en classes des individus (= SN) ayant des caractéristiques communes. Elle peut donc permettre d'accéder aisément à toutes les données contenant de l'information. En plus, la représentation de la classification tire parti de l'héritage des descriptions entre classes.

Notre objectif était donc de définir une approche de classification qui permette non seulement de construire mais aussi de raffiner des classifications. Ce processus est de toute évidence un processus itératif et automatisé. Au terme des différents problèmes présentés, nous définissons le processus de la façon suivante :

Processus de la classification pour l'organisation :

- *Etant donné :*
  - un ensemble d'individus SN et leurs descriptions associées (leurs structurations)
  - des connaissances sur la structure classificatoire recherchée
  - des connaissances pour évaluer la qualité d'une classification
- *Trouver une classification des individus SN, c'est à dire :*
  - un ensemble de classes qui regroupent ces individus
  - une définition intensionnelle de chacune de ces classes
  - une organisation hiérarchique de ces classes

Une telle problématique demande une nouvelle approche de la classification qui s'éloigne des familles classiques (CLUSTER, COBWEB, …).
La mise en oeuvre d'une telle approche requiert un processus d'itération en se basant sur les connaissances explicites fournies par l'analyseur morpho-syntaxique : $R_{elation\_1}$(N, SN), $R_{elation\_2}$ (SN, SN+, SN–).

Telle est précisément la principale difficulté du raisonnement par classification, de même que la mise en évidence des liens dans le regroupement conceptuel incrémental (la formation des concepts, des classes et des liens) [RAJMAN, 98a-b].

- **Exemple 1 :**

Nous extrayons un échantillon de résultats sur la classification représentant les relations $R_{elation\_1}$(N, SN), $R_{elation\_2}$ (SN, SN+, SN–) autour du thème sur les « robots » :

N	SN
robot	▪ les **robots** ▪ les **robots** industriels
univers	▪ le univers ▪ un univers ▪ le univers de les **robots** ▪ tout le univers de les **robots** ▪ le univers de la intelligence artificielle à les **robots** industriels
intelligence	▪ la intelligence artificielle à les **robots** industriels
boite à musique	▪ les boites à musique à les **robots**
...	▪ ...

Une représentation graphique (*Fig.VI.6.3a.*) de ces deux relations, avec le logiciel AVRIL v.3/95 (CNRS-IRPEACS), permet de montrer le treillis constitué entre les divers éléments de connaissances textuelles.

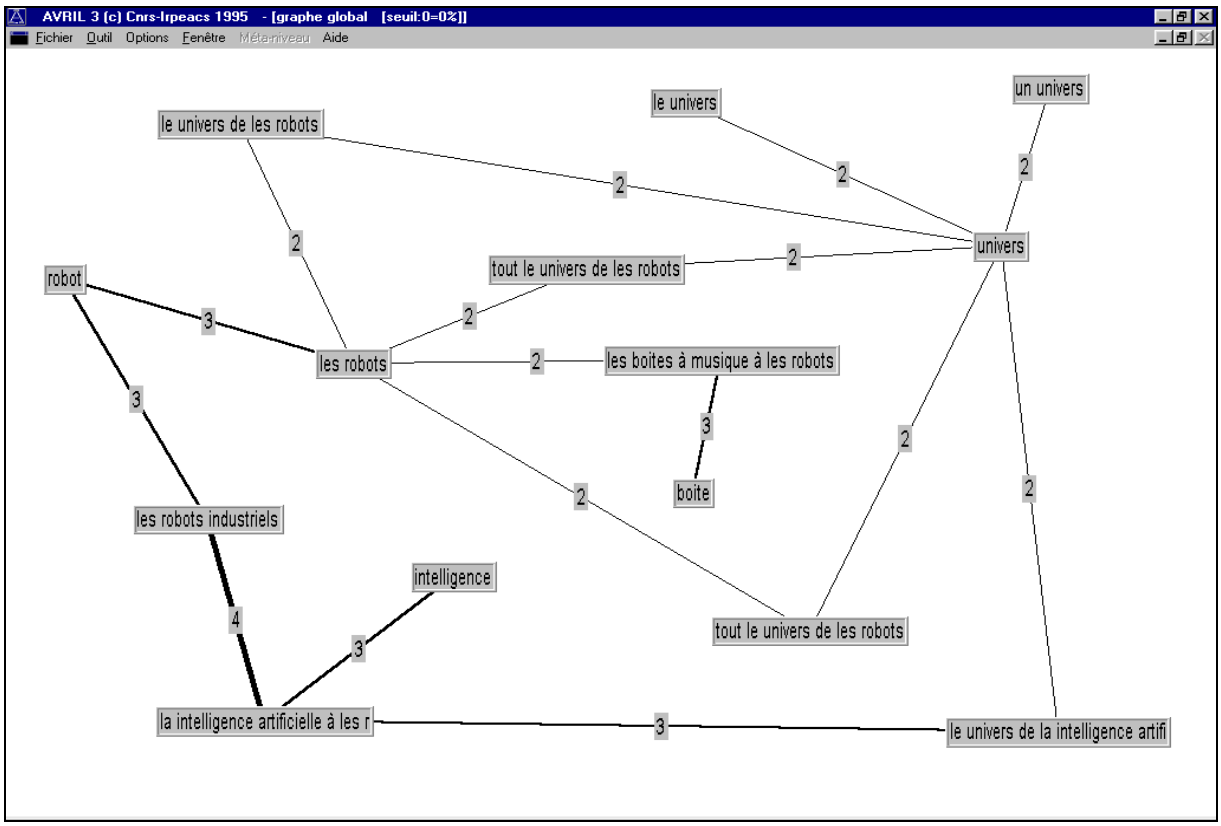


Fig.VI.6.3a. Sous graphe du réseau des connaissances autour du thème « robot ».

- **Exemple 2 :**

Un second échantillon de résultats sur la classification représentant les relations $R_{elation\_1}$(N, SN), $R_{elation\_2}$ (SN, SN+, SN–) autour du thème sur la « guerre » (*Fig.VI.6.3b.*).

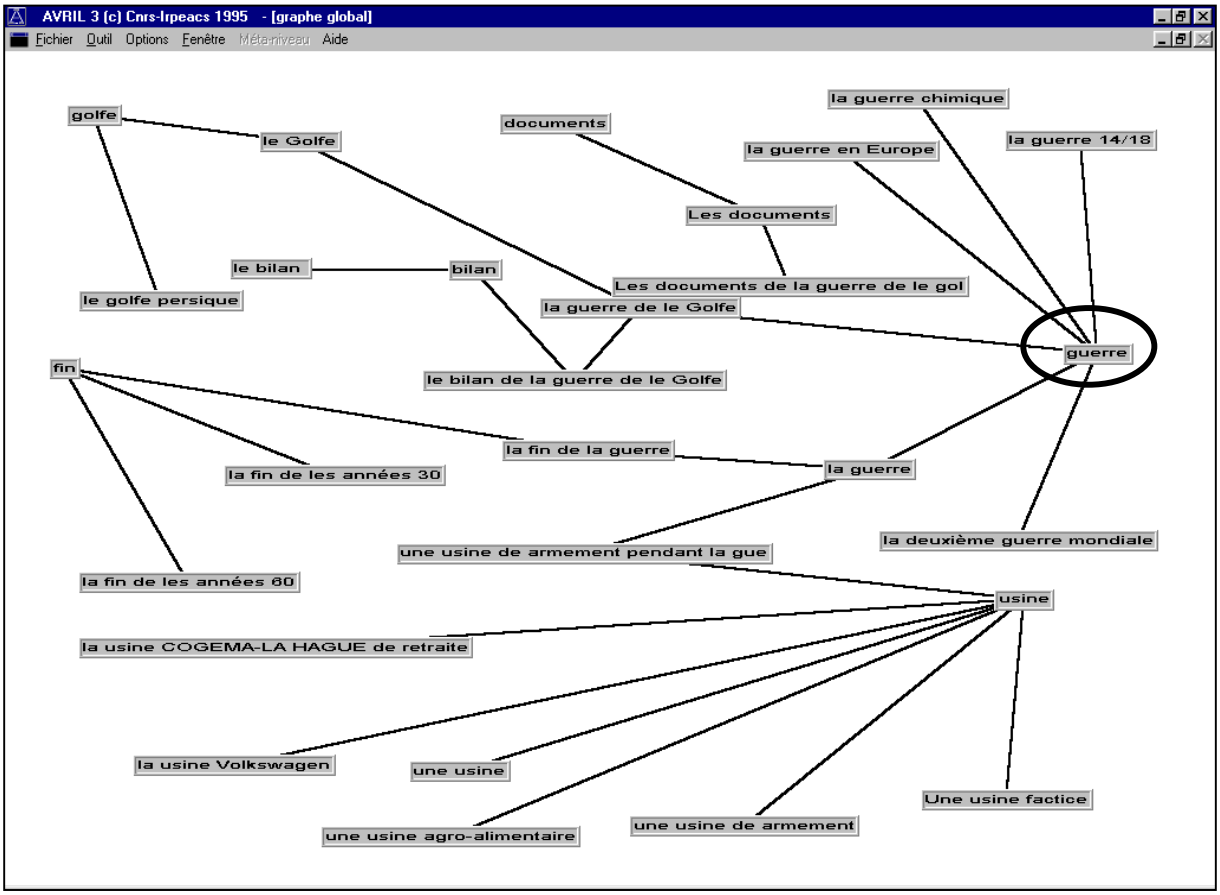


Fig.VI.6.3b. Sous graphe du réseau des connaissances autour du thème « guerre ».

VI.6.4- Evaluation logico-sémantique

Le modèle établi par les précurseurs du groupe SYDO, pour définir la morphologie générale du descripteur en indexation documentaire, permet de mettre en valeur des formes linguistiques spécifiques : le syntagme nominal et sa complexité constituent la forme privilégiée.

Ce modèle permet également de déterminer des niveaux d'appréhension différents du descripteur (indexeur, utilisateur). En dernier, le modèle développe de façon concise la problématique du descripteur comme unité de discours.

L'approche logico-sémantique du syntagme nominal, dans laquelle M. Le Guern (1991) [LE GUERN, 91] et M. Amar (1997) [AMAR, 97] développent le statut du descripteur et les fondements théoriques de l'indexation, fait apparaître la morphologie canonique du descripteur qui est celle du syntagme nominal et de ses propriétés interprétatives constitutives.

Cette approche privilégie le point de vue de la recherche documentaire sur l'aspect de la capture de descripteurs du point de vue de l'utilisateur. Le modèle permet de spécifier la morphologie du descripteur dans le contexte de l'indexation explicative : rôle assigné aux descripteurs dans le système d'information.

➢ Comparaison entre les descripteurs de l'INA et les SN

chapeau et résumé (par 1 notice) :

	nbr PHR	nbr SN	SN/PHR
chapeau	2	18,3	9,15
résumé	5,8	35,4	6,1
Analyse automatique	**7,8**	**53,7**	**15,25**

	nbr PHR	nbr DES	DES/PHR
INA			
Analyse INA	**7,8**	**13**	**1,66**

- **Illustration sur une notice INA et de ces résumés analysés :**

- Notice constituée par l'INA (rédaction du résumé) :

Numéro DL DL T 19950107 M6 007
Titre Airbus Industrie
résumé **AIRBUS INDUSTRIE** parallèlement à l'Europe qui est en pleine mutation (chute du mur de Berlin, Tunnel sous la Manche...) prépare son avenir avec le **A 321**.
Le **A 300 B** fit son vol inaugural, le 28 octobre 1972 avec Bernard ZIEGLER comme pilote d'essai. Mais toute la période de **construction** fut semé d'embauches tant économiques, politiques que techniques. La **construction** de la voilure fut construite à Chester en Angleterre, la section principale du fuselage à Hambourg en Allemagne, le cockpit et la section centrale à Saint-Nazaire et à Toulon par l'Aérospatial. Le plus dur fut d'acheminer les différentes parties à BLagnac pour l'assemblage du prototype. Les voies aériennes et routières fut empruntées non sans mal. Les tests draconiens du programme de certification sont passés avec réussite.
Le succès commercial de l'**Airbus A300B** a permis au consortium européen d'élargir sa gamme d'appareils.
Suivant les mêmes techniques de recherche, de **construction** et d'assemblage, **AIRBUS INDUSTRIE** construira les modèles **A310**, **A320**, **A330** et **A340**.

- Analyse INA (descripteurs du thésaurus) :

Numéro DL DL T 19950107 M6 007
Titre Airbus Industrie
Descripteurs construction aéronautique; Airbus Industrie; avion à réaction (A300, A310, A320, A330)

- Analyse AUTOMATIQUE (SN extraits automatiquement du résumé) :

Numéro DL DL T 19950107 M6 007
Titre Airbus Industrie

N, SN, SN+, SN-	N	SN	SN+	SN–
	…			
	A 300	Le A 300	_	_
	A 321	le A 321	son avenir avec le A 321	_
	A300	A300 B	_	_
	Airbus	le Airbus A300B	Le succès commercial de le Airbus A300B	_
	ans	20 ans de AIRBUS	_	_

N, SN, SN+, SN-	N	SN	SN+	SN–
	avenir	son avenir avec le A 321	_	le A 321
	construction	La construction de la voilure	_	la voilure
	fuselage	le fuselage à Hambourg	la section principale de le fuselage à Hambourg	Hambourg
	voilure	la voilure	La construction de la voilure	_
	…			

Dans cet exemple, nous avons choisi seulement les N et SN (extraits automatiques) qui se rapprochent des termes descripteurs INA. Nous observons non seulement la richesse morphologique et syntaxique des SN, mais également leur pouvoir interprétatif pour rendre compte des connaissances dans le discours (texte résumé).
Cet avantage logico-sémantique dans l'indexation par les SN permet de mettre en valeur les formes linguistiques spécifiques du SN dans l'ordre de ses propriétés interprétatives constitutives et le contexte de son discours.
Un avantage à plusieurs niveaux (morphologique, syntaxique et interprétatif du SN) est nettement observable, qui dépasse le terme descripteur (DES) par thésaurus INA. On peut bien étendre cet avantage devant une indexation par mots-clés uni-termes ou multi-termes.

VI.7- Conclusion

Par la confrontation systématique entre modes de représentation de la langue, les niveaux d'analyse et les implications pratiques, nous parvenons à distinguer, d'une part, le fonctionnement de la langue, et d'autre part, son utilisation en particulier dans le contexte de l'indexation.
Dès lors, nous pouvons rapporter aux propriétés de la langue un point de vue expérimentale portant sur l'indexation automatique.

Selon ce point de vue, nous avons montré que :

- L'indexation repose sur l'exploitation des propriétés inhérentes à la linguistique : la signification lexicale, la portée référentielle entre SN–objets de l'univers, et la construction de la référence dans le discours. « *Cet ensemble de propriétés linguistique fonde la possibilité de l'indexation…* » [AMAR, 97], p.366.

- La spécificité de l'indexation automatique se fonde dans la construction des thème de discours. Un tel projet se donne comme objectif la reconnaissance des objets de discours par une matérialisation des formes linguistiques spécifiques : le SN.

- Le rôle du descripteur s'exprime clairement d'abord sur plusieurs points :
  - logique,
  - linguistique,
  - classificatoire,
  - logico-sémantique.

Ensuite, il s'exprime également comme une double relation de référence :

- en premier à l'appartenance à une classe d'objet,
- en second, à la désignation d'un objet spécifique dans cette classe,
- en troisième, à l'exploitation des liens entre références (du spécifique au générique et vice-versa) et
- en dernier, à la manipulation des choix d'accès à l'information, du mode d'emploi (relation prédicat intensionnel *versus* objet extensionnel, relation de synonymie entre prédicats intensionnels).

Cet aspect de notre recherche a été conduit dans le but de faire émerger les propriétés linguistiques et le traitement du langage dans une pratique expérimentale sur l'indexation automatique documentaire. L'étendue du projet a montré la nécessité de coordonner d'autres sources et de stratégies dans l'exploration, le raisonnement, et l'exploitation des objets de discours spécifiques à l'indexation et à la recherche d'informations.

Les problèmes qui restent à résoudre et qui permettront à nos propositions de trouver une voie de réalisation complète et effective sont nombreux. En grande partie, les extensions prometteuses de ce travail, après des constructions fondées sur des aspects logique, linguistique et logico-sémantique de l'indexation, se retrouvent dans le regroupement conceptuel pour l'organisation des connaissances. Ce domaine est en constante réactivité de l'intelligence artificielle et plus précisément « l'extraction de connaissances à partir des données symboliques ou textuelles ».

Conclusion

> Au terme de ce travail de recherche, nous prendrons appui sur l'analyse de L. Bloomfield[1] (1933) qui soulignait particulièrement que : " *Ce qui concerne le sens est le point faible des études sur le langage, et le restera jusqu'à ce que nos connaissances aient avancé bien loin de leur état actuel* ".
>
> En réalité, nos connaissances sur le sujet ont-elles vraiment avancé depuis cette époque ? Nous nous trouvons tout juste au début du III^{ème} millénaire où les travaux dans le domaine du traitement automatique de la langue naturelle continuent à contribuer à cet éventuel avancement.

Pourtant, la langue est vue comme un canal qui véhicule du sens, ce qui n'empêche pas à poser de sentencieux problèmes pour les traitements automatiques. Ces derniers passent systématiquement par le biais d'une représentation formelle. Une telle représentation doit posséder suffisamment de propriétés lui permettant de rendre compte de ces diverses compétences linguistiques.

Notre travail a consisté à poser en partie le problème de la syntaxe comme guide d'un processus d'interprétation.

Dans le même ordre d'idées, B. Bachimont[2] (1992) soutenait que " *la connaissance est un état mental sémantiquement interprété. Quand les connaissances sont objectivées, elles prennent la forme de représentations symboliques et syntaxiques, véhiculant la signification que possèdent les connaissances. Passer d'une représentation à une connaissance, c'est effectuer une interprétation* ".

Autour de ces réflexions, nous avons pu construire nos hypothèses de travail sur la base du traitement automatique du langage vers une représentation morpho-syntaxique, et l'interprétation de celle-ci comme une représentation de la connaissance vers son organisation logico-sémantique. Et c'est ainsi que nous avons fait le passage de l'écrit vers la gestion des connaissances qui a pour but l'indexation et la recherche d'information.
Si notre approche présente des intérêts, il est indéniable qu'elle possède également des limites.
Pour toutes ces raisons, nous allons d'abord présenter dans ce qui suit, un résumé sur la contribution de notre travail, les limites de la Plate-forme d'analyse morpho-syntaxique dédiée à l'indexation et la recherche d'information et, enfin, les perspectives que nous envisageons pour ce travail.

---

[1] : Citation de Bloomfield Leonard (en 1933, *Language*, New York Press) et reproduite par Bobrow Daniel 1968, in Natural language input for a computer problem solving system, *Semantic information processing*, MIT Press, Cambridge.

[2] : Bachimont Bruno 1992, Le contrôle dans les systèmes à base de connaissances, Hermès, Paris.

1. Contribution de la thèse

La contribution de ce travail de thèse s'inscrit au sein d'un domaine multidisciplinaire regroupant le traitement automatique du langage naturel, l'indexation dans un système d'information documentaire et l'organisation des connaissances autour de l'information écrite. Sa particularité consiste en la mise à disposition d'outils pour le traitement automatique de l'information.

A ce titre, nous avons, dans un premier temps, clarifié l'Espace de Recherche dans lequel nous nous sommes situés. Nous avons tout d'abord posé les bases de notre discussion sur la connaissance écrite, sa nature, son origine et ses contributions dans la transmission des savoirs. Nous avons précisé les apports de cette connaissance qui ne se limite pas à l'écrit mais étendue à l'audio-visuel. L'idée de cette réflexion nous a amené à intégrer un nouveau composant concernant l'étude même de cet objet « *connaissance* » sur corpus. Le corpus concerné est un ensemble de notices documentaires de l'Institut National de l'Audiovisuel à Paris (INA France) puis étendu à des sources sur le Web.
Dans ces notices, nous avons retrouvé différents types de résumés de contenu selon le type du document : texte, image (fixe ou animée), audio ou audiovisuel/vidéo. Les résumés de l'INA ont été développés en fonction d'une grille d'analyse de contenu de l'audiovisuel. On peut relever dans cette grille d'analyse plusieurs types d'informations :

- la position des plans et la description plan par plan,
- les mouvements de caméra,
- les personnes identifiées, les lieux de l'action et,
- la distinction entre image et son.

Le fruit de notre réflexion sur l'étude du corpus consistant à l'analyse des résumés a donné naissance à un modèle « cognitif » de rédaction. D'après l'analyse statistique appliquée sur ses résumés, les résultats obtenus ont montré une stabilité grammaticale dans les constructions de la phrase. Cette révélation d'ordre grammaticale cache en réalité une stabilité de rédaction des textes résumés que nous exploitons au profit de la grammaire de réécriture de notre analyseur morpho-syntaxique.

Le mécanisme d'analyse automatique des résumés s'est concrétisé par la conception d'un noyau d'indexation automatique. Le noyau d'indexation se scinde en trois parties :

- la réalisation des différents outils automatiques servant à l'analyse morphologique du langage naturel,
- l'implémentation de l'analyseur morpho-syntaxique basé sur les règles de réécriture dans le modèle de rédaction de la phrase, et
- l'extraction des syntagmes nominaux à partir des arbres syntagmatiques de la phrase analysée. L'architecture du noyau d'indexation est basée sur les automates à transitions augmentées en cascade (CATN) de W. Woods.

Cette dernière étape permet, d'une part, l'organisation, le stockage dans une base de données relationnelles, les constructions logico-sémantiques des syntagmes nominaux et, d'autre part, l'organisation des connaissances autour du syntagme nominal.

Dans notre recherche, cet aspect sur l'organisation des connaissances a été conduit dans le but de faire émerger les propriétés linguistiques et le traitement du langage dans une pratique expérimentale sur l'indexation automatique documentaire. Nous avons montré la nécessité de coordonner d'autres sources et stratégies dans l'exploration de ces propriétés. Il s'agit du mode de raisonnement et de la technique d'exploitation des objets du discours spécifiques à la gestion des connaissances (comme étape préalable à la recherche d'information).
Ces deux derniers aspects (mode et technique) intégrés dans le processus de la présentation et de l'organisation du syntagme nominal offrent des scénarii pertinents pour la recherche d'informations.

Cette Plate-forme d'analyse dans sa logique de fonctionnement sert comme un outil d'investigation orienté vers l'organisation et la gestion des connaissances écrites.

2. Limites de la Plate-forme d'analyse

L'intérêt que l'on porte aux langages documentaires et particulièrement aux processus d'indexation et de recherche d'information provient notamment du fait de leur brusque prolifération, au sens des approches et méthodologies déployées. L'élaboration d'un système d'information documentaire, par l'appel aux nouvelles approches ou modèles hybrides d'indexation, a donné sens aux réflexions sur les problèmes liés à l'indexation et la recherche d'information : l'objet de la connaissance.

Ce domaine de réflexion pluridisciplinaire, nous l'avons abordé sur le plan théorique par le biais de la linguistique computationnelle. Ce contexte privilégié offre l'analyse et la description du langage naturel et son intégration dans un processus automatisé. Cet avantage consolide notre idée du départ sur la connexion du processus du traitement du langage à celui de la recherche d'information. Cette connexion se matérialise par l'intermédiaire d'un processus d'organisation des connaissances autour du syntagme nominal.

Au sens de N.I. Žinkin (1962) et de R. Jacobson (1963), *les langages documentaires sont des métalangages qui sont tous autant que le langage-objet, un aspect de notre comportement verbal* : comme tel, ils constituent un problème linguistique. Nos moyens d'approche seront ceux basés sur des études linguistiques.

En fait, le mécanisme dans sa conception est bien adapté à la problématique. Celle-ci étant liée au statut du descripteur dans un processus d'indexation. Le recours à la linguistique est un procédé fiable dans le passage des formes textuelles au codage recherché, autrement dit, les procédures linguistiques permettent d'introduire une rigueur auto-suffisante pour catégoriser, regrouper et interpréter les mots liés à leur forme de surface : les connaissances.

La difficulté principale qui surgit dans la conception de ce mécanisme et son adaptation aux objectifs fixés, est de savoir quelle méthodologie de développement il serait souhaitable d'adopter pour un tel système ?
Avant d'envisager l'approche pour la réalisation de ce système, un critère comme l'intégrabilité est nécessaire à considérer.

L'intégrabilité consiste à pouvoir ajouter aisément des éléments (outils) à un système et la possibilité d'en étendre aisément leurs fonctionnalités.

Pour la programmation linguistique nous n'avons pas employé un langage spécialisé, mais nous avons adapté un modèle linguistique à un formalisme de programmation puis son implémentation dans un langage de type objet (C++). Ce langage a la facilité d'intégrer aisément des routines procédurales et des interfaçages avec une base de données relationnelle.

Une seconde difficulté qui s'est imposé dans la programmation linguistique, est que l'implémentation d'outils linguistiques nécessitent des composants de nature variée. Il est souvent difficile de catégoriser clairement les composants d'un système de TALN lié à un SRI. On ne pourra cependant préciser ces composants qu'après avoir caractérisé les objets, les actions, et les acteurs qui interviennent dans un tel système.

A l'heure actuelle nous observons une limite liée à quelques tâches de conception pure. Pour la limitation linguistique, nous citons par exemple le traitement de la coordination et le traitement des déterminants complexes du français.
Quant à la limitation pour la recherche d'information, nous citons par exemple l'incomplétude du dictionnaire des synonymes et la représentation graphique du treillis des connaissances autour du syntagme nominal.

3. Perspectives

Dans une certaine mesure de clarté, il était question de présenter les définitions relatives au modèle de reconnaissance, d'analyse et de représentation du syntagme nominal. Nous avons également exposé les acceptions inhérentes à la définition de la classification. Et enfin, nous avons proposé une approche de construction automatique de classifications conceptuelles. Cette approche prend en compte les sorties de l'analyseur morpho-syntaxique que sont le syntagme nominal, son emboîtement dans une structure hiérarchique et son centre nominal.

L'outil d'analyse en question, c'est-à-dire l'analyseur morpho-syntaxique, pourra servir comme aide à la rédaction de textes, et pourra produire de manière automatique leur indexation.

Dans une démarche classificatoire explicite, le syntagme nominal est la mise en oeuvre de deux organisations logiques différentes. Au niveau N, on a l'intervention de la logique intensionnelle qui est une « *logique sans référentiel et sans classe, constituée de relations et de propriétés envisagées indépendamment de quelque objet que ce soit*. Par contre au niveau N'' (et en partie au niveau N'), celui-ci relève de la logique extensionnelle et on peut envisager une classe d'objets avec *la mise en relation des mots et des choses (objets)*.

A l'issue de ces deux organisations logiques, des associations « naturelles » apparaissent reliant les références au trait d'une classe d'objets. La première association concerne les références en relation avec d'autres : la relation du type SN_i-SN_j. La seconde concerne les références en relation avec le trait de sa classe d'objets : relation du type N_i–SN_i.

Ainsi, on découvre l'organisation des différentes parties du discours au moyen d'une construction classificatoire. La classification en question est élaborée selon une organisation

conceptuelle du SN. Celui-ci permet d'établir un réseau cohérent d'informations structurées et de construire de manière interactive un ordre opératoire à partir de la description de ses objets.

Le premier type de perspective est lié à l'amélioration de ce système d'organisation. Il s'agit en particulier de synchroniser le module d'extraction des syntagmes nominaux au module de construction de l'Espace des classifications de façon à permettre une utilisation complète des fonctionnalités du système à partir de l'interface.

Une autre perspective dans le développement de l'Espace de classification consiste à modifier la construction de l'association des références. Nous l'avons exprimée par la relation du type SN-SN. Dans cette relation, il s'agit d'un double lien entre $SN_i \rightarrow SN_{i+1}$ et $SN_i \rightarrow SN_{i-1}$. L'avantage de cette relation consiste, d'une part, à dépasser la limite qui fixe le maximum d'emboîtement (niveau i : $SN_i \rightarrow SN_{i-1/i+1}$) dans un syntagme nominal complexe et, d'autre part, à chercher dans une référence SN_k d'autres références spécifiques (respectivement génériques) de niveau n/ $n \leq k$ (respectivement $n \geq k$).
Quant à la recherche des références de niveau $p > k$ (respectivement $p<k$), nous disposons seulement de $p = k+1$ (respectivement $p=k-1$). Nous espérons augmenter les références génériques/spécifiques qui dépassent le niveau k-1/k+1 selon un mode interactif.

Ceci répondrait à une demande de recherche d'informations qui conditionne l'observation de la référence des relations génériques et spécifiques sur plusieurs niveaux.

Enfin, pour que notre système soit un véritable outil d'investigation dans les bases de données textuelles, il est nécessaire d'intégrer un mécanisme de gestion de classification avec mémorisation des étapes de recherche. Un outil permettant de visualiser ce processus et qui offre la possibilité de revenir à une étape antérieure en ayant sa place dans les outils de recherche.

L'enjeu de cette perspective de recherche est un véritable traitement, dans la construction des classifications avec une mémoire, et de descriptions structurées de l'Espace de recherche d'informations. Une des voies de recherche qui nous semble intéressante vise à trouver des mécanismes permettant de prendre en compte les appariements entre références cherchées, références obtenues et classification conceptuelle de celles-ci. Une telle approche réside dans l'apprentissage de concepts.

ANNEXE :

Plate-forme d'analyse du Langage Naturel, d'Indexation Automatique et de Recherche d'Information

A. Introduction

La Plate-forme d'Analyse du langage naturel, d'Indexation automatique et de Recherche d'information est développée sous le système Windows®.

Dans la réalisation de cette Plate-forme d'Analyse, il s'agit de coordonner la conception de divers outils dont :

- **SIMBAD :** Système d'Indexation du Multimédia Basé sur l'Analyse de contenu Documentaire, regroupant l'Analyseur morpho-syntaxique et le Noyau d'indexation
- **Dic-Fr :** le dictionnaire électronique du français,
- **SRI-Fr :** Système de Recherche d'Information basé sur les connaissances autour du syntagme nominal, et
- **Syn-Fr :** le dictionnaire électronique des synonymes du français.

Nous avons également exploité d'autres outils externes au profit de notre Plate-forme d'analyse pour :

- la visualisation des arbres syntagmatiques à la sortie de l'analyseur morpho-syntaxique, et
- la visualisation du treillis des connaissances autour du syntagme nominal.

Le côté expérimental de notre travail de recherche nous a amené à évoluer vers un système complet d'ingénierie linguistique et de la gestion des connaissances. Ces deux aspects caractérisent le traitement automatique de l'information écrite (langue française).
La première étape consistait à se servir des travaux théoriques du Groupe de Recherche SYDO-Lyon comme support caractérisant nos outils, les implémentant puis les faisant évoluer. Les bases algorithmiques ont été formulées et optimisées dans un environnement de programmations procédurales et objets (MS.™ VISUAL C/C++).

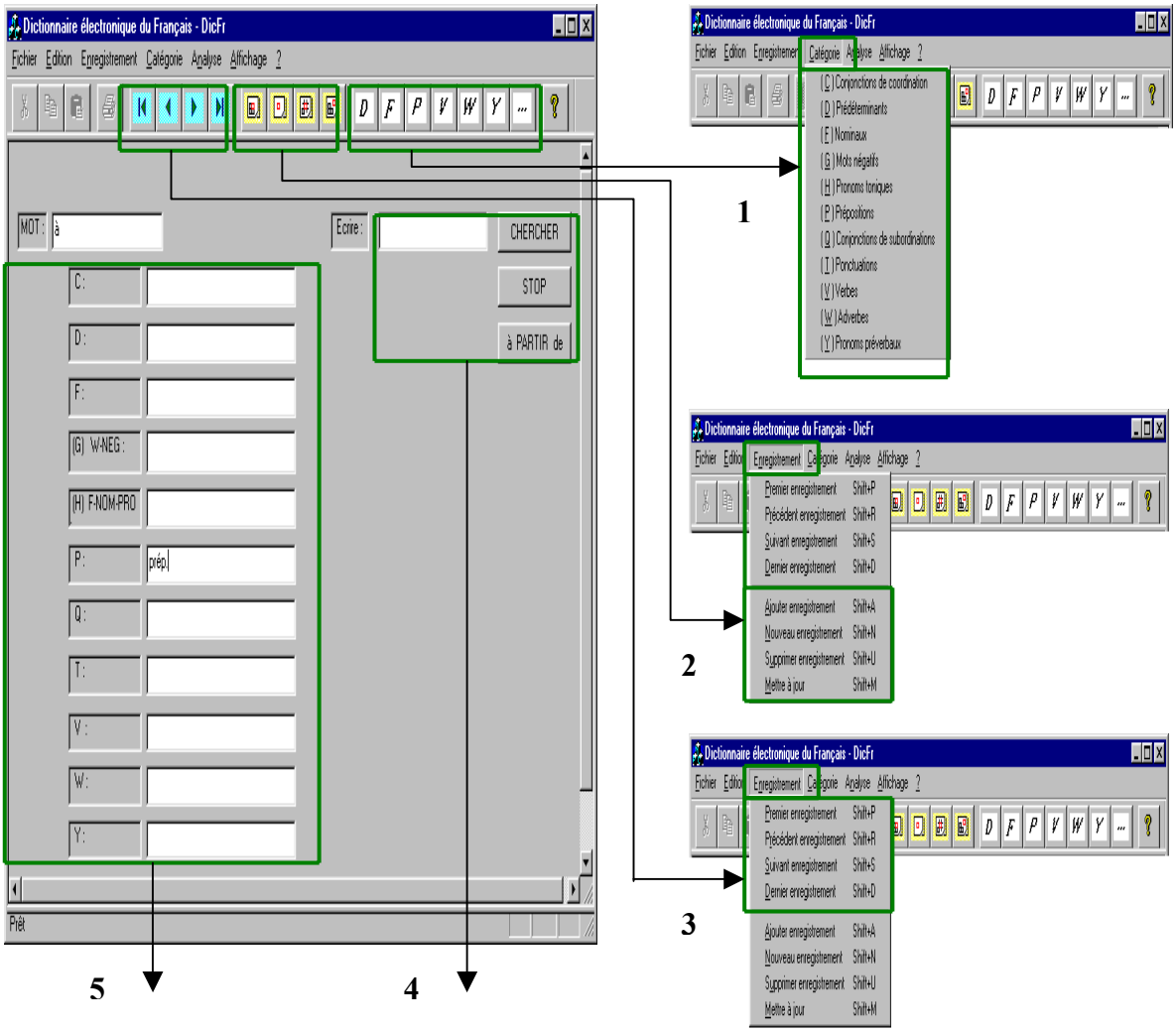
Le second objectif a consisté à coordonner le fonctionnement des outils expérimentaux permettant d'observer les résultats théoriques. Des incrémentations théoriques et pratiques étaient nécessaires lors de cette étape pour une mise en oeuvre cohérente.

En dernier, nous avons consolidé nos outils par des interfaces de dialogue Homme-machine. Cela nous facilite la manipulation en nous permettant d'observer les résultats d'analyse et leur cohérence. Un cadre de vérification et de tests pour faire évoluer notre Plate-forme a ainsi été proposé.
Dans ce qui suit, nous présenterons les fonctionnalités des outils réalisés pour la Plate-forme d'Analyse.

A.1.- Interface graphique du dictionnaire électronique du français

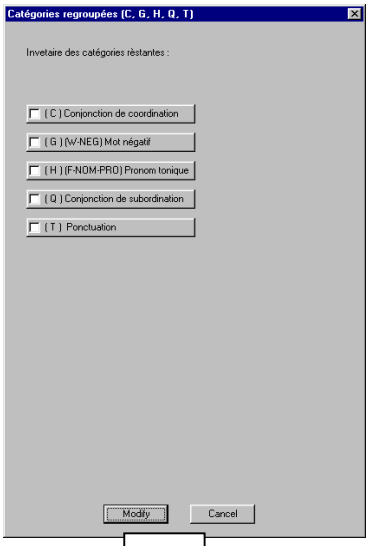
A.1.1- Le dictionnaire et ses fonctionnalités

Le dictionnaire électronique du français était le premier outil développé pour la Plate-forme. Il est basé sur le Modèle Classificatoire SYDO dans la représentation du lexique : [→5], afin de faciliter l'analyse ou la recherche d'un mot (forme) : [→4]. Plusieurs fonctionnalités dans les menus **Enregistrement** et **Catégorie** du dictionnaire sont offertes, comme les commandes de saisie : [→1], d'ajout ou de mise à jour : [→2] d'un mot selon une catégorie. Aussi bien, pour les commandes de déplacements par mot ou groupe de mots dans le dictionnaire : [→3].

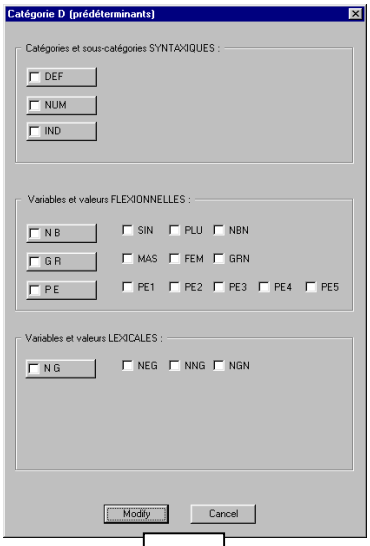


A.1.2- Formatage du lexique selon le modèle classificatoire SYDO

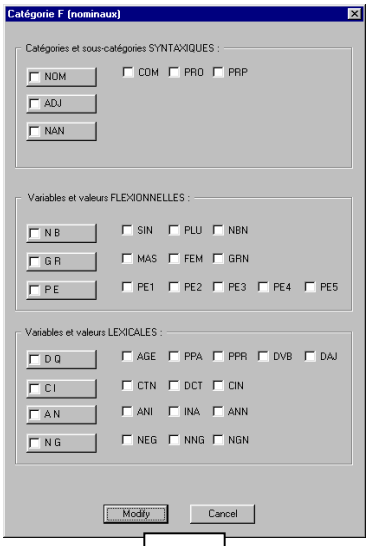
Dans le menu **Catégorie,** plusieurs commandes sont disponibles par catégorie syntaxique. La sélection d'une commande fait appel à une **Boîte de Dialogue** {[1],...,[7]} représentant une catégorie {C,D,F,P,Q,T,V,W,Y}. Dans chaque Boîte de Dialogue, trois classes de définition sont représentées : syntaxique, flexionnelle et lexicale (cf. *Tab.V.3.2.2*). Pour chaque définition d'une classe, des boutons de sélection sont attribués aux variables et valeurs spécifiques à cette dernière. Pour une nouvelle forme, il suffit de saisir, sélectionner et valider.


Catégories regroupées (C, G, H, Q, T)
Invetaire des catégories restantes :
(C) Conjonction de coordination
(G) (W-NEG) Mot négatif
(H) (F-NOM-PRO) Pronom tonique
(Q) Conjonction de subordination
(T) Ponctuation
Modify
Cancel

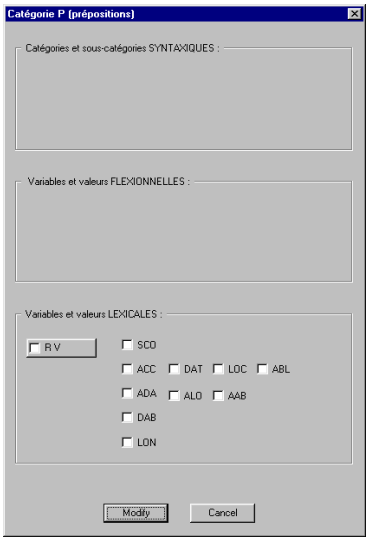
1


Catégorie D (prédéterminants)
Catégories et sous-catégories SYNTAXIQUES :
DEF
NUM
IND
Variables et valeurs FLEXIONNELLES :
N B
SIN PLU NBN
G R
MAS FEM GRN
P E
PE1 PE2 PE3 PE4 PE5
Variables et valeurs LEXICALES :
N G
NEG NNG NGN
Modify
Cancel

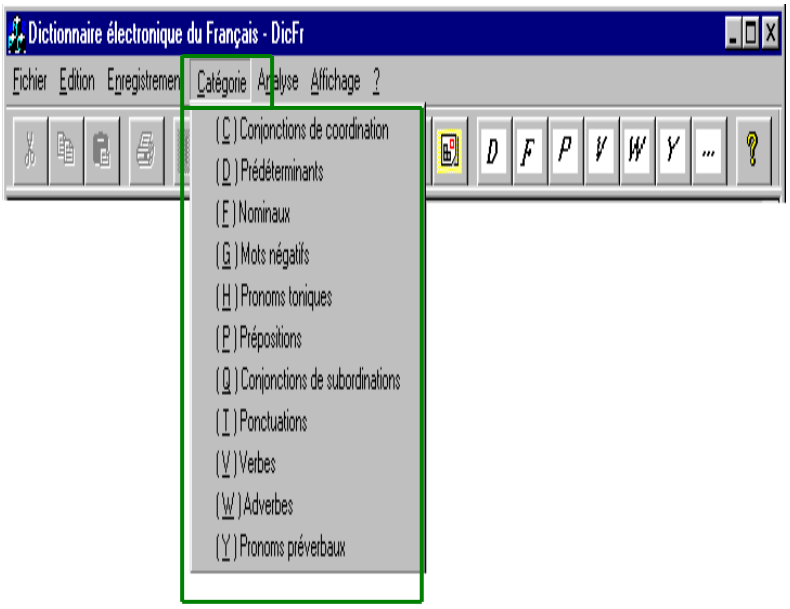
2

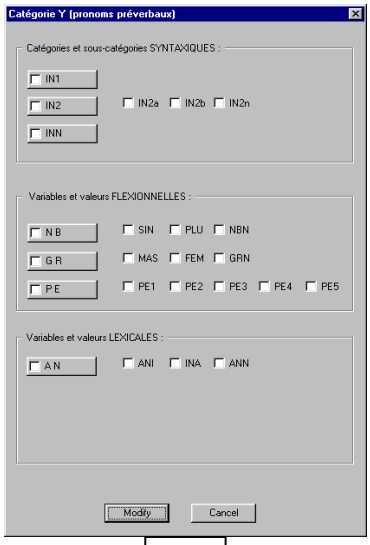

Catégorie F (nominaux)
Catégories et sous-catégories SYNTAXIQUES :
NOM
COM PRO PRP
ADJ
NAN
Variables et valeurs FLEXIONNELLES :
N B
SIN PLU NBN
G R
MAS FEM GRN
P E
PE1 PE2 PE3 PE4 PE5
Variables et valeurs LEXICALES :
D Q
AGE PPA PPR DVB DAJ
C I
CTN DCT CIN
A N
ANI INA ANN
N G
NEG NNG NGN
Modify
Cancel

3

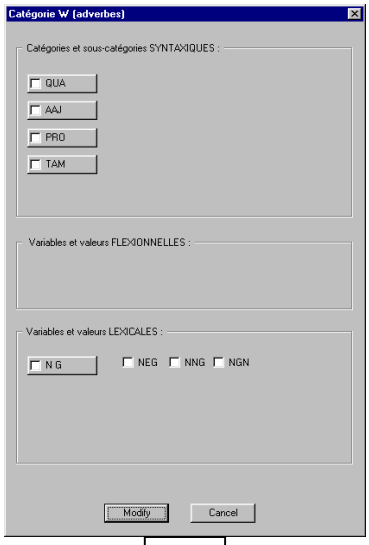

Catégorie P (prépositions)
Catégories et sous-catégories SYNTAXIQUES :
Variables et valeurs FLEXIONNELLES :
Variables et valeurs LEXICALES :
R V
SCO
ACC DAT LOC ABL
ADA ALO AAB
DAB
LON
Modify
Cancel

4

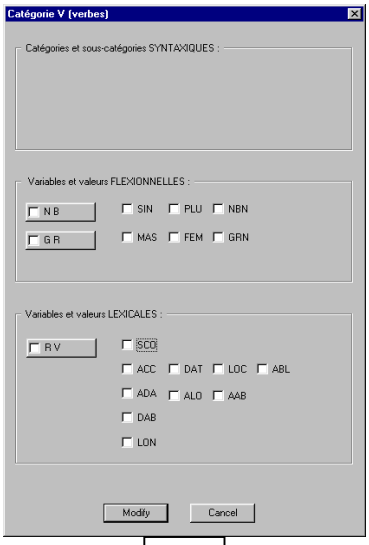

Dictionnaire électronique du Français - DicFr
Fichier Edition Enregistremen Catégorie Analyse Affichage ?
(C) Conjonctions de coordination
(D) Prédéterminants
(F) Nominaux
(G) Mots négatifs
(H) Pronoms toniques
(P) Prépositions
(Q) Conjonctions de subordinations
(T) Ponctuations
(V) Verbes
(W) Adverbes
(Y) Pronoms préverbaux


Catégorie Y (pronoms préverbaux)
Catégories et sous-catégories SYNTAXIQUES :
IN1
IN2
IN2a IN2b IN2n
INN
Variables et valeurs FLEXIONNELLES :
N B
SIN PLU NBN
G R
MAS FEM GRN
P E
PE1 PE2 PE3 PE4 PE5
Variables et valeurs LEXICALES :
A N
ANI INA ANN
Modify
Cancel

5


Catégorie W (adverbes)
Catégories et sous-catégories SYNTAXIQUES :
QUA
AAJ
PRO
TAM
Variables et valeurs FLEXIONNELLES :
Variables et valeurs LEXICALES :
N G
NEG NNG NGN
Modify
Cancel

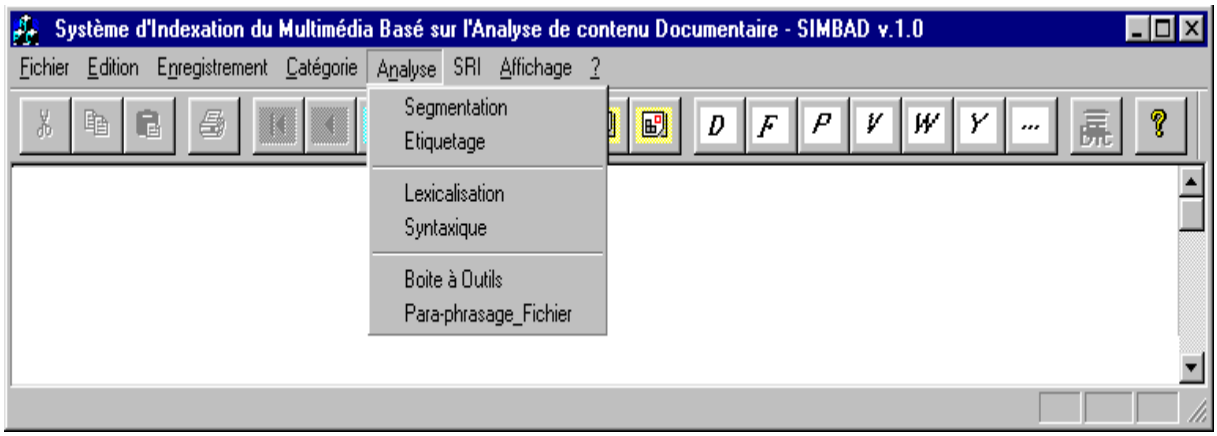
6


Catégorie V (verbes)
Catégories et sous-catégories SYNTAXIQUES :
Variables et valeurs FLEXIONNELLES :
N B
SIN PLU NBN
G R
MAS FEM GRN
Variables et valeurs LEXICALES :
R V
SCO
ACC DAT LOC ABL
ADA ALO AAB
DAB
LON
Modify
Cancel

7

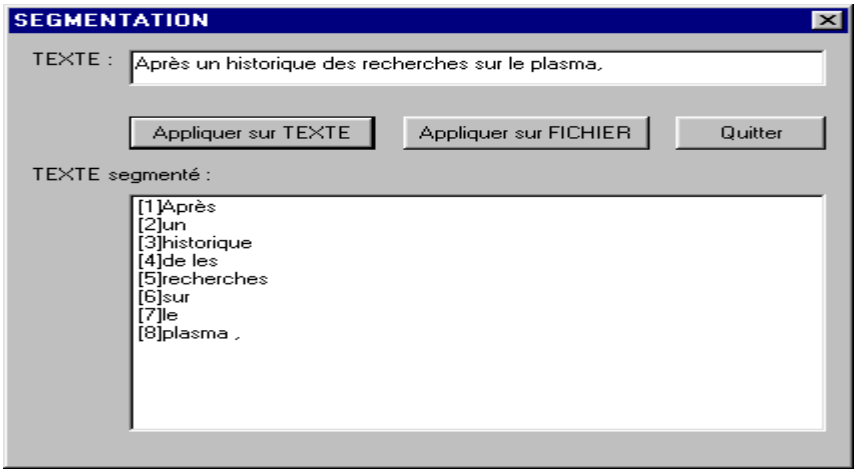
A.2.- Interface graphique de l'analyseur morpho-syntaxique

L'analyse morpho-syntaxique consiste essentiellement à mettre en oeuvre un processus sur plusieurs étapes d'analyse. Essentiellement, après l'analyse des formes dans la surface d'un texte, le processus cherchera à dégager un maximum d'informations pour permettre une structuration du texte par des regroupements d'unités syntagmatiques.



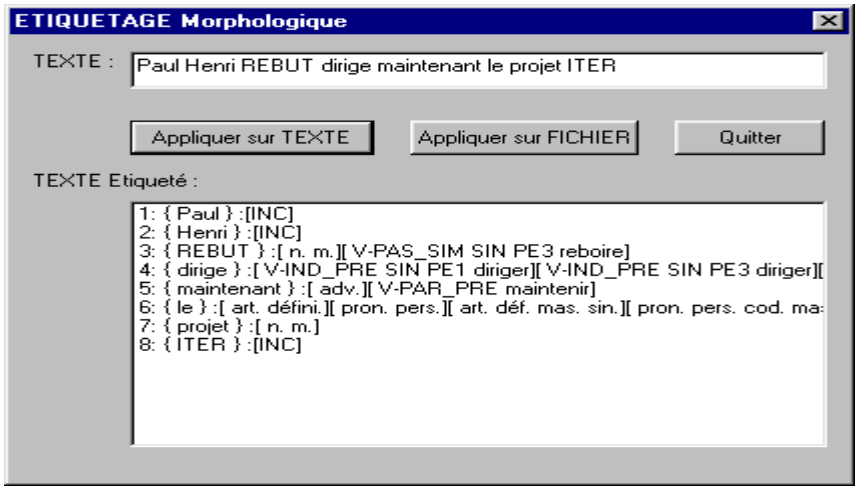
Dans le menu **Analyse**, plusieurs commandes sont disponibles regroupant des modules (étapes) d'analyse séparés puis interconnectés : prétraitement morpho-syntaxique, traitement morphologique, segmentation textuelle, résolution des ambiguïtés et l'analyse morpho-syntaxique.

A.2.1- Segmentation : régularisation morpho-syntaxique de surface



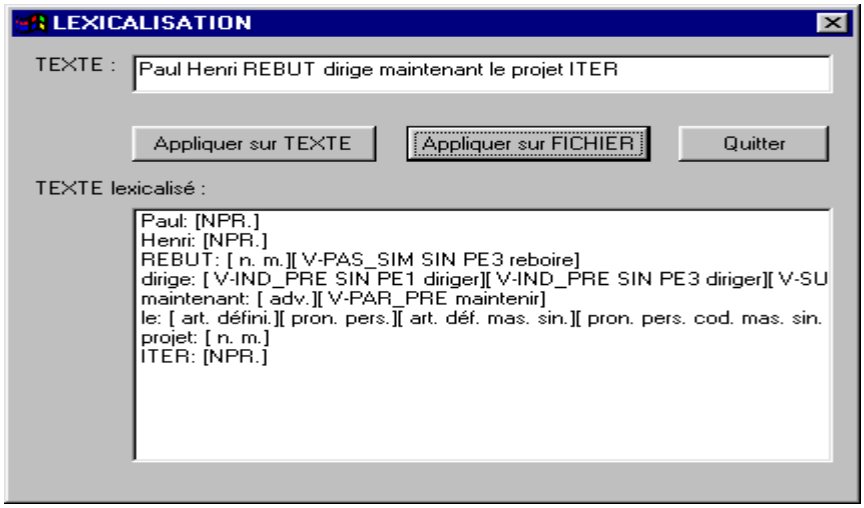
Le modèle classificatoire n'est applicable que si la séquence des formes textuelles est soumise à une pré-analyse. Cette étape a pour but d'éliminer de la surface d'un texte toutes les formes résultant d'un amalgame (amalgames orthographiques et morphologiques, mots en /qu-/). Ainsi, nous réduisons le nombre des mots d'un texte à classer dans le Modèle Classificatoire.

A.2.2- Analyse morphologique



A l'issu de la régularisation morpho-syntaxique de la surface, les mots d'un texte sont complétés par les informations du dictionnaire : traits relatifs aux classes, variables et valeurs. Certains mots restent inconnus [INC.] et qui seront de nouveau analyser.

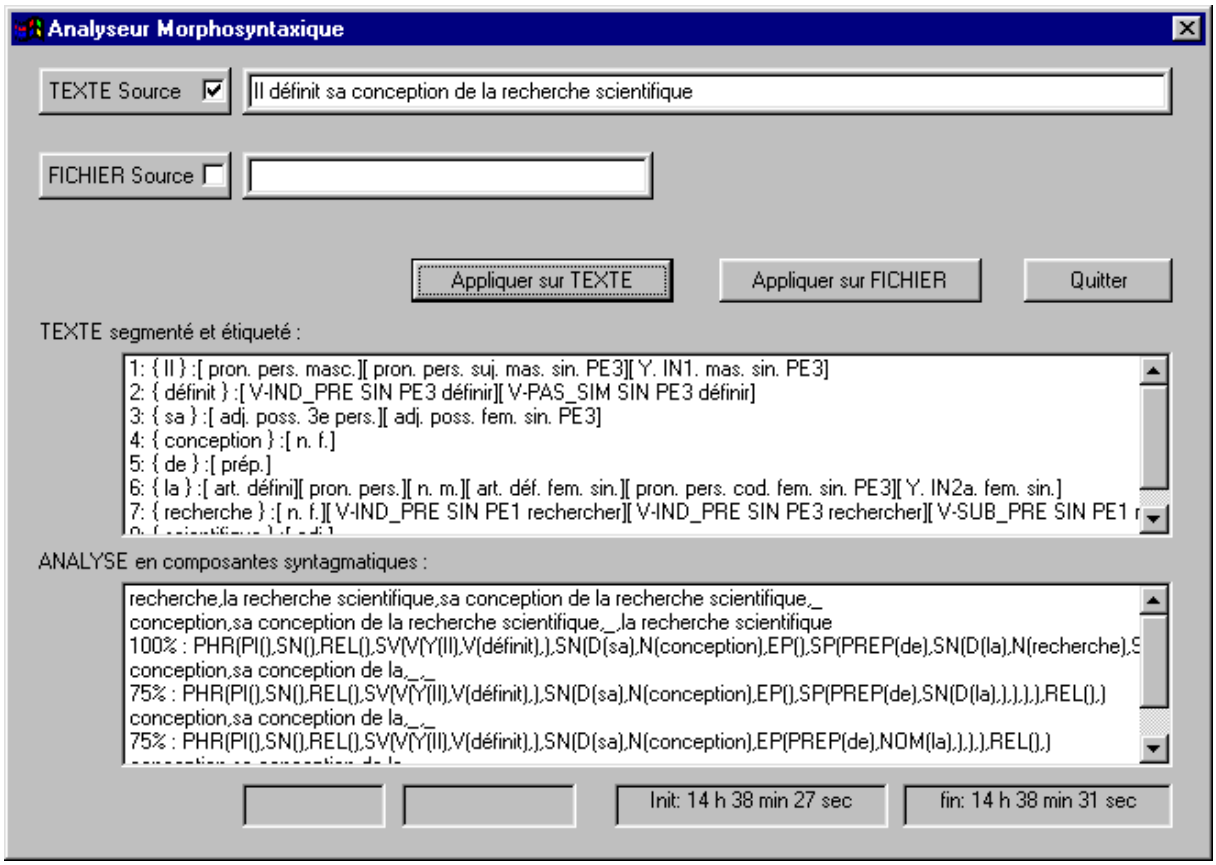
A.2.3- Analyse lexicale : levée partielle des ambiguïtés morphologiques



Les mots inconnus sont de nouveaux analysés par appariement de leur forme ou de leur orthographe, comme les noms propres, etc.
D'autres seront résolus par rapport à leur emplacement dans la surface textuelle ou la catégorie des mots adjacents, cas des mots inexistants dans le dictionnaire.
Cette analyse consiste à attribuer au mot ses propriétés grammaticales (catégorie, sous-catégorie, et valeurs de variables) afin de pouvoir engager l'analyse syntaxique proprement dite.

A.2.4- Analyse morpho-syntaxique

- Application sur un texte saisi directement :

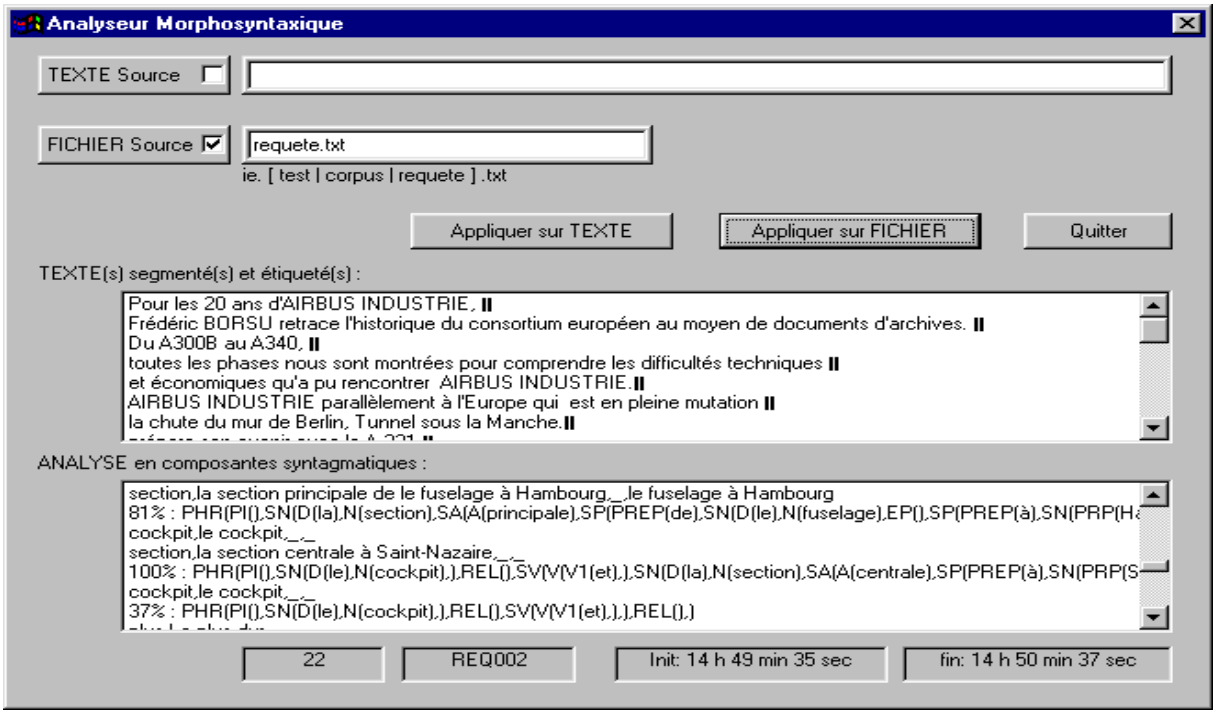


La phrase (ou un segment de phrase) est écrite dans le contrôle de saisie **[TEXTE source]** puis par demande d'analyse automatique, sur le bouton **[Appliquer sur TEXTE]**, produira les résultats d'analyse jusqu'à la production des structures syntagmatiques (arbres d'analyse). L'outil d'indexation continuera le processus par la lecture de ces structures et l'extraction des connaissances : N, SN, SN+, SN- en faisant référence aux notices, requêtes ou phrases. A la fin de ce processus, le fichier inverse est constitué.

Si l'analyse de la phrase est ambiguë plusieurs analyses sont produites avec leurs coefficients d'analyse. Ce coefficient représente les mots de la phrase ayant contribués à la production des structures syntagmatiques et qui ont été validés par l'analyseur morpho-syntaxique.

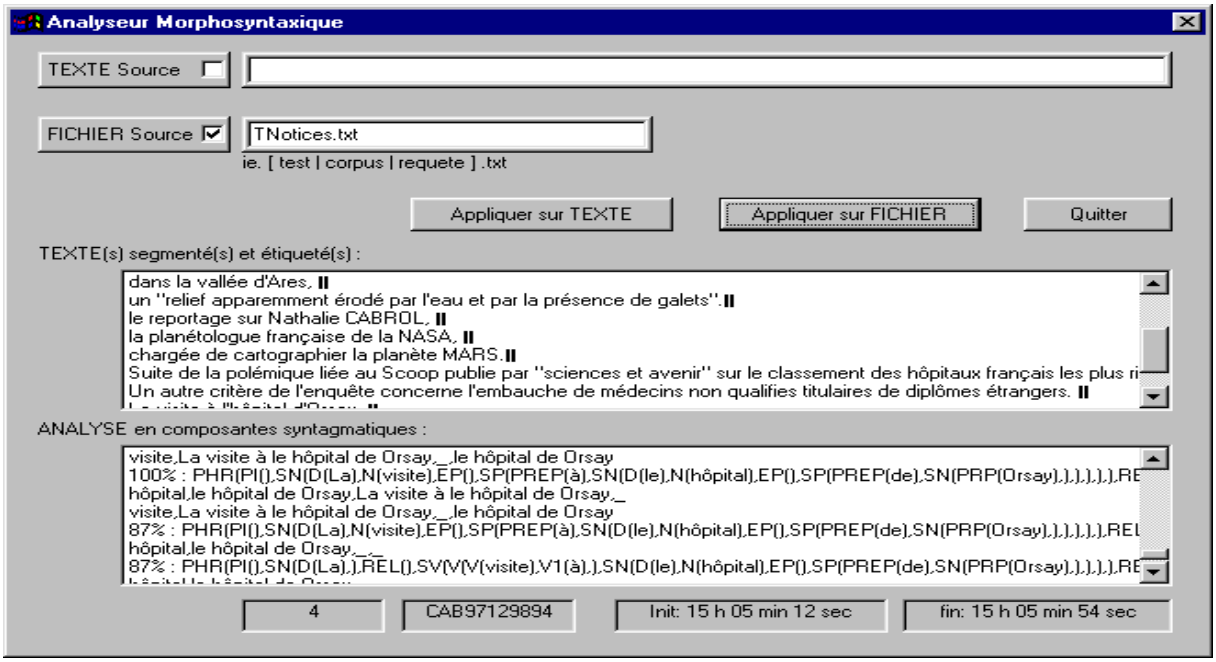
Ce coefficient est également calculé pour une requête d'interrogation en langage naturel.

- Application sur un fichier textuel : requête d'interrogation



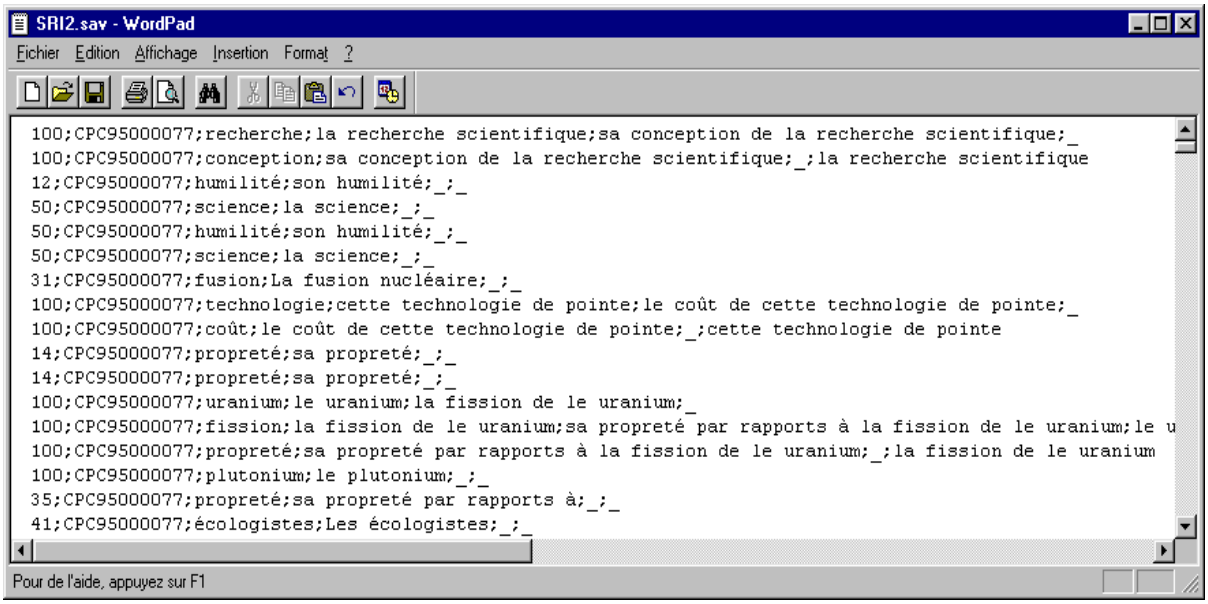
Une requête est exprimée sous formes de paragraphes : ensemble de phrases qui constituent les thèmes du sujet de recherche. L'analyse automatique de la requête produira les mêmes structures d'information que celles obtenues lors de l'analyse d'une phrase.

- Application sur un fichier textuel : résumés de documents



Idem. pour l'analyse d'un fichier contenant des résumés textuels. Initialement, ces résumés sont issus du corpus des notices documentaires. Le fichier inverse est produit à la fin de l'analyse totale du corpus. Celui-ci constituera notre base de connaissances.

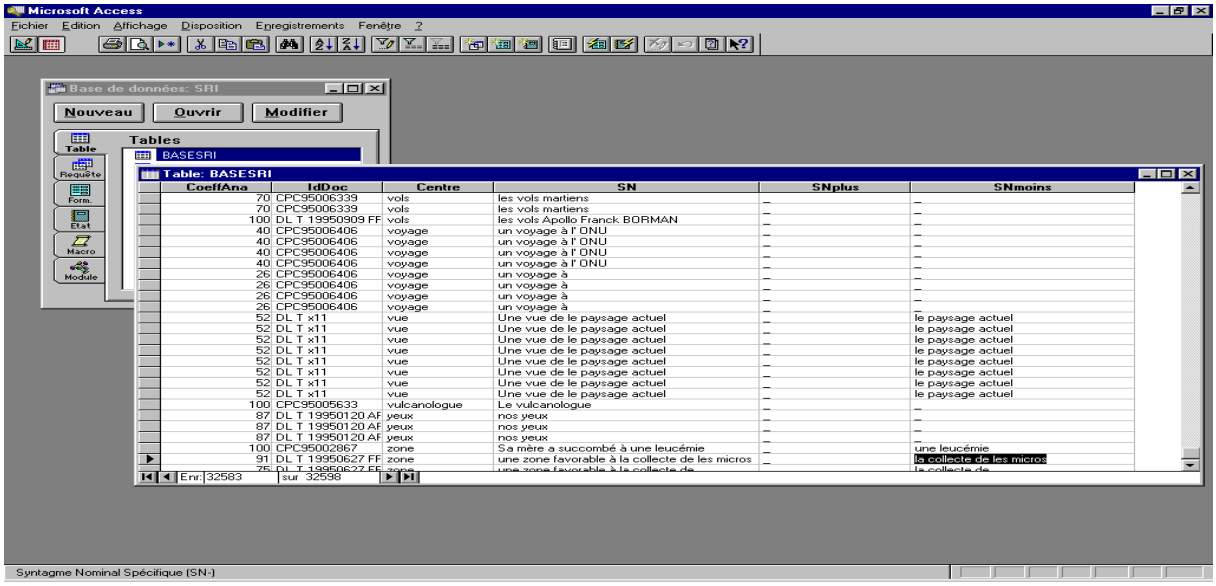
A.2.5- Extraction automatique des connaissances autour du syntagme nominal : Génération du Fichier inverse



Le fichier inverse est constitué lors de l'analyse d'un contenu textuel. Il mémorise les différents éléments informationnels sur le document et sur les connaissances extraites de son contenu :

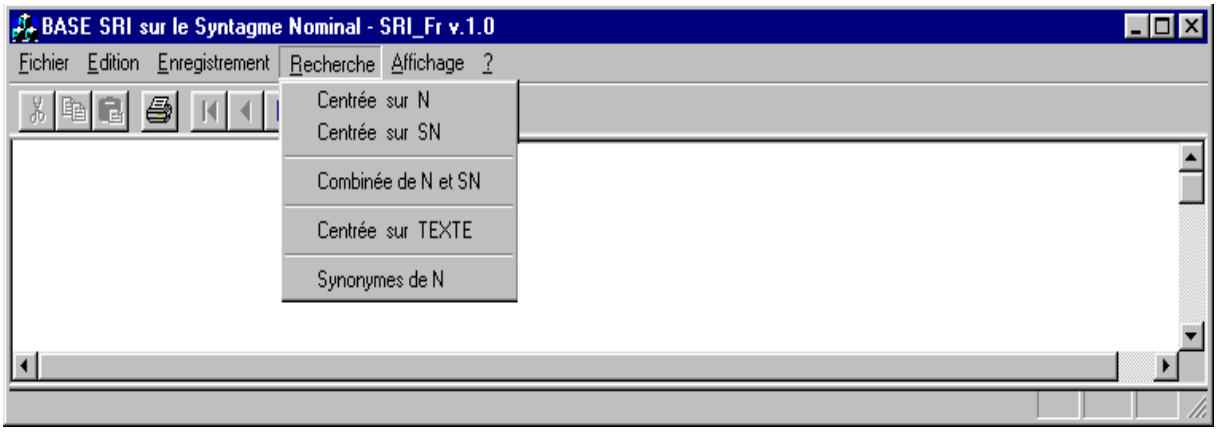
- le coefficient d'analyse dans chaque phrase ayant aboutie à l'extraction des SN : *0..100 %*
- l'identificateur du document analysé : *ID_document,*
- le centre Nominal (centre du syntagme SN_i) : N_i ,
- le syntagme nominal et ses emboîtements : $SN_i, SN_{i+1}^{+}, SN_{i-1}^{-}$.

A.2.6- Fichier inverse : intégration en Base de Données relationnelles



A la suite de l'analyse automatique du corpus, le fichier des connaissances extraites est intégré dans une base de données. Cette base permettra une exploitation optimale des sources de connaissances et comme *Fichier Inverse* dans le processus de recherche d'informations.

A.3- Interface graphique du système de recherche d'information



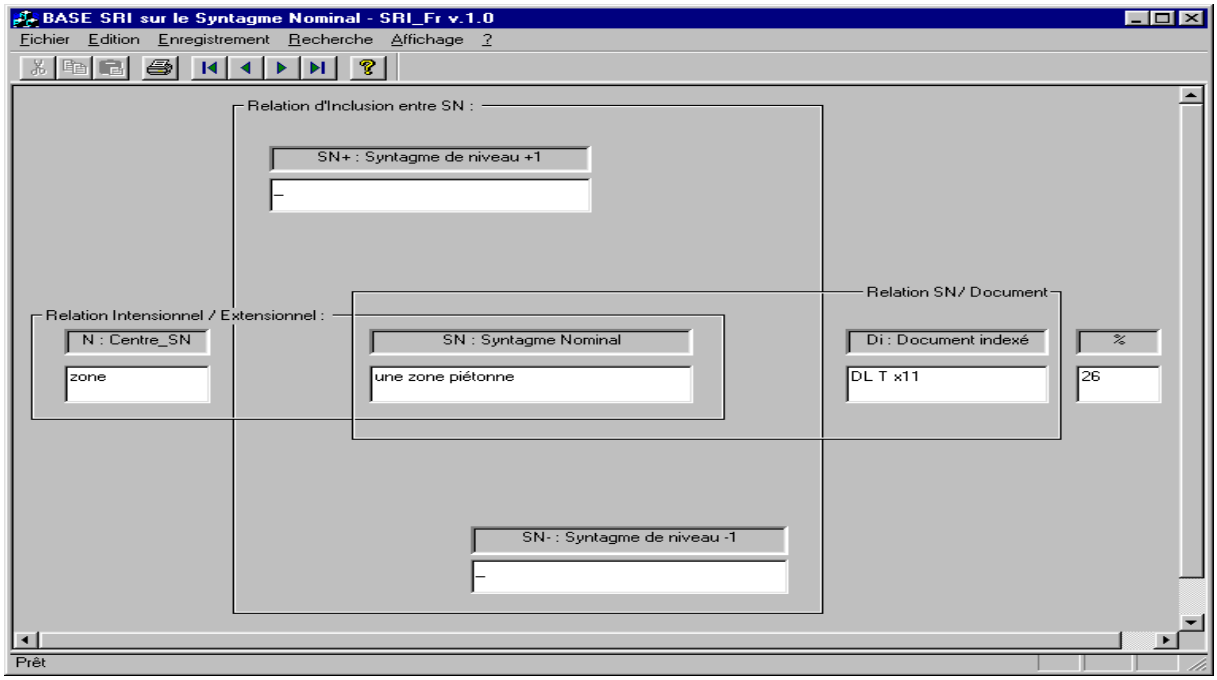
L'interface graphique du système de recherche d'information présente les fonctionnalités inhérentes aux connaissances liées à la gestion des syntagmes nominaux :

– Recherche centrée sur le centre nominal (N),
– Recherche centrée sur le syntagme nominal (SN),
– Recherche combinée entre le centre nominal (N) et le syntagme nominal (SN),
– Recherche centrée sur une requête textuelle (ensemble de phrases),
– Recherche opérant sur les synonymes (N_{syn}) d'un centre nominal N.

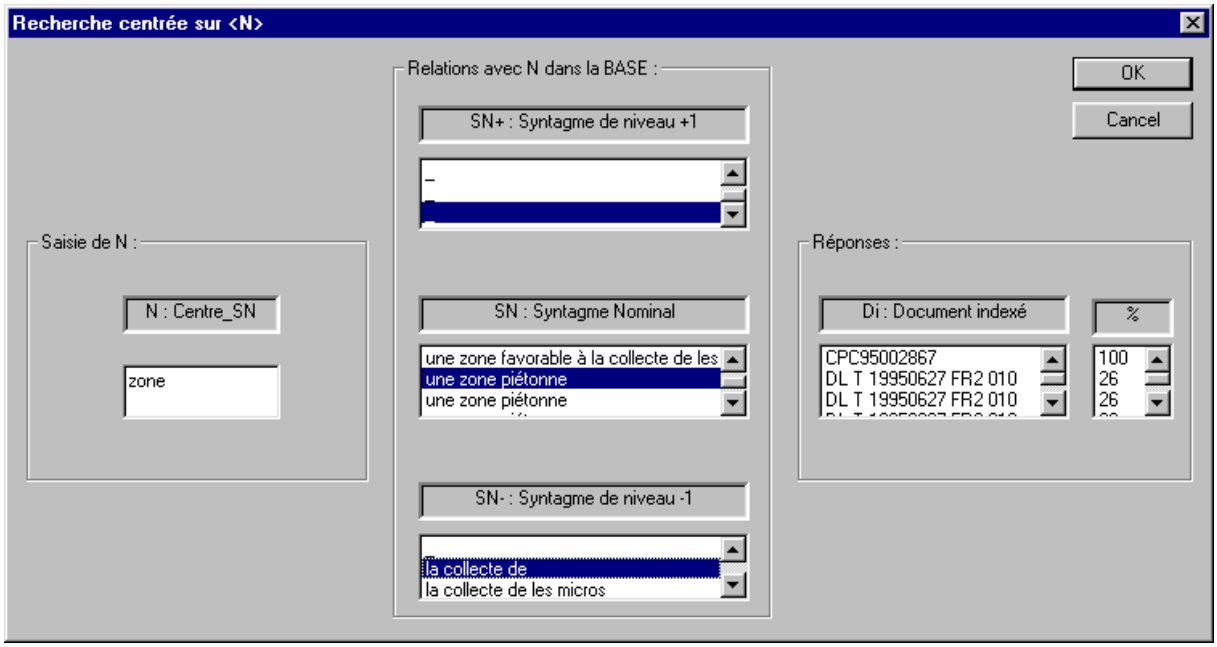
Ce dernier cas opérant sur les synonymes est présenté pour remédier à l'échecs d'une recherche centrée sur N ou celle centrée sur SN.

A.3.1- Consultation de la base index-documents : Fichier inverse

Le fichier inverse constitué, d'une part, de la référence aux documents indexés et, d'autre part, des connaissances autour du SN, est muni d'une interface graphique de consultation.

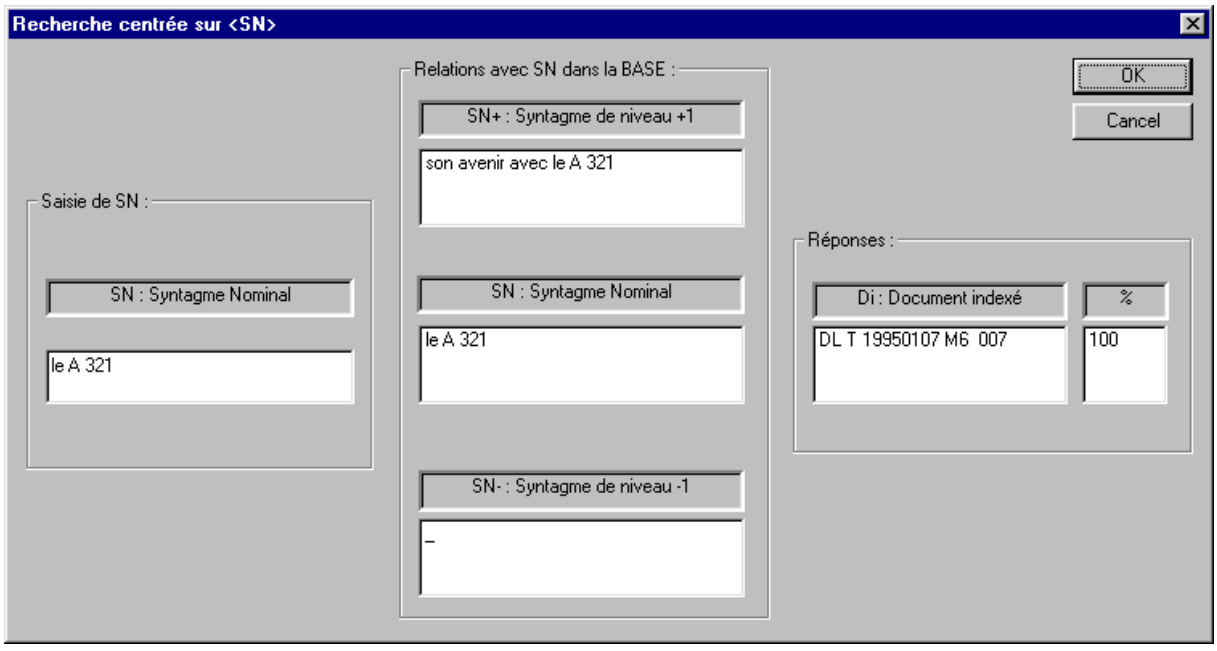


A.3.2- Recherche basée sur le centre nominal



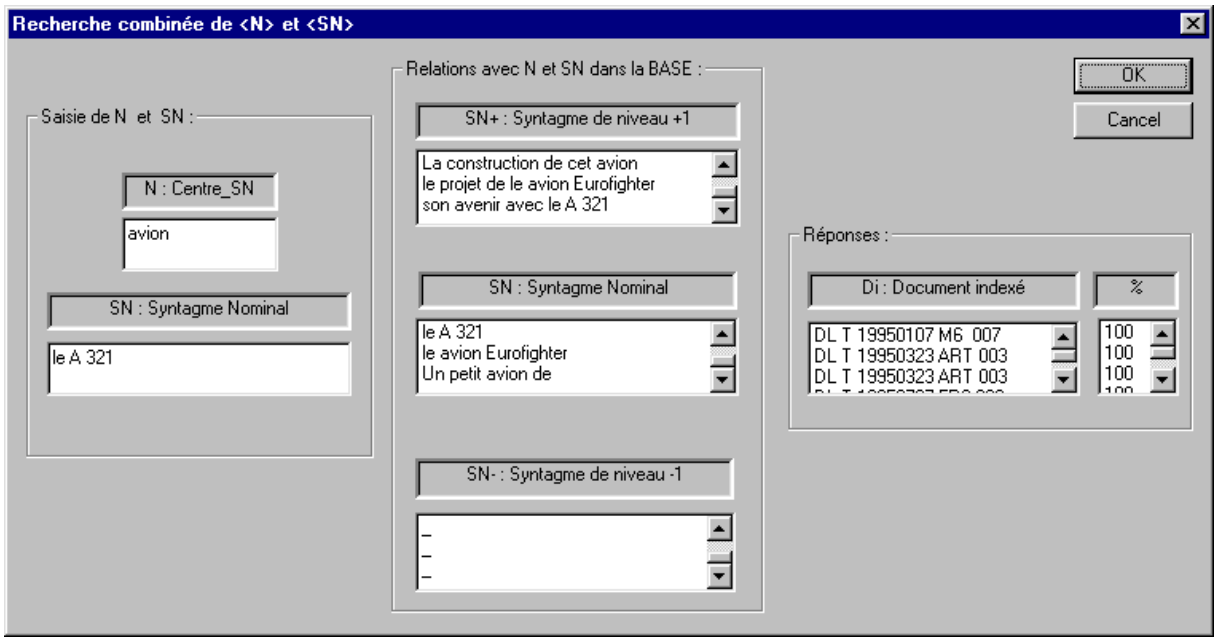
L'interface graphique de la recherche centrée sur un centre nominal (N) consiste dans son mode opératoire à saisir un prédicat libre N puis d'appliquer la recherche par [OK]. Les informations relatives à ce prédicat sont directement affichées en faisant référence aux relations (N–SN) de la base avec les documents associés.

A.3.3- Recherche basée sur le syntagme nominal



L'interface graphique de la recherche centrée sur un syntagme nominal (SN) consiste dans son mode opératoire à saisir la chaîne référentielle (prédicat lié saturé ou non) SN puis d'appliquer la recherche par [OK]. Les informations relatives au SN sont directement affichées en faisant référence aux relations (SN–SN+), (SN–SN$^-$) de la base avec les documents associés.

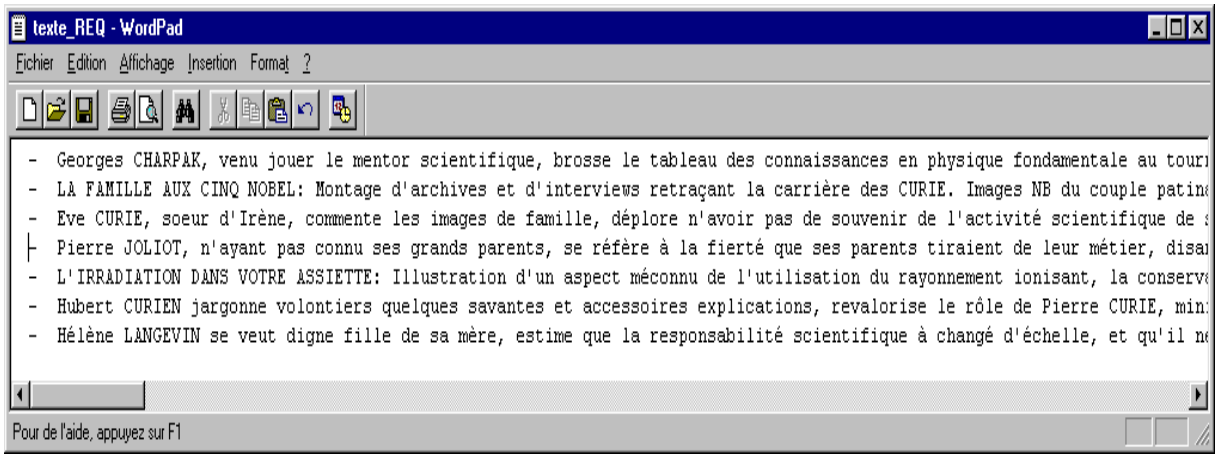
A.3.4- Recherche combinée entre centre et syntagme nominal



L'interface graphique de la recherche combinée entre un centre nominal (Ni) et un syntagme nominal (SNj) consiste dans son mode opératoire à saisir les deux chaînes Ni et SNj puis d'appliquer la recherche par [OK]. Les informations relatives aux Ni et SNj sont directement affichées en faisant référence aux relations (N_i–SN_i) et celles (SN_j–SN_j+) et (SN_j–SN_j^-) de la base avec les documents associés.

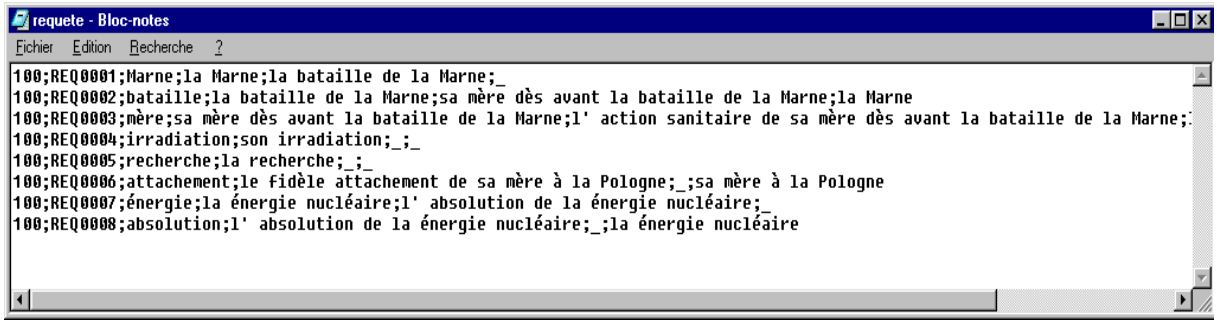
A.3.5- Recherche basée sur une requête exprimée en langage naturel

- Descriptif textuel :



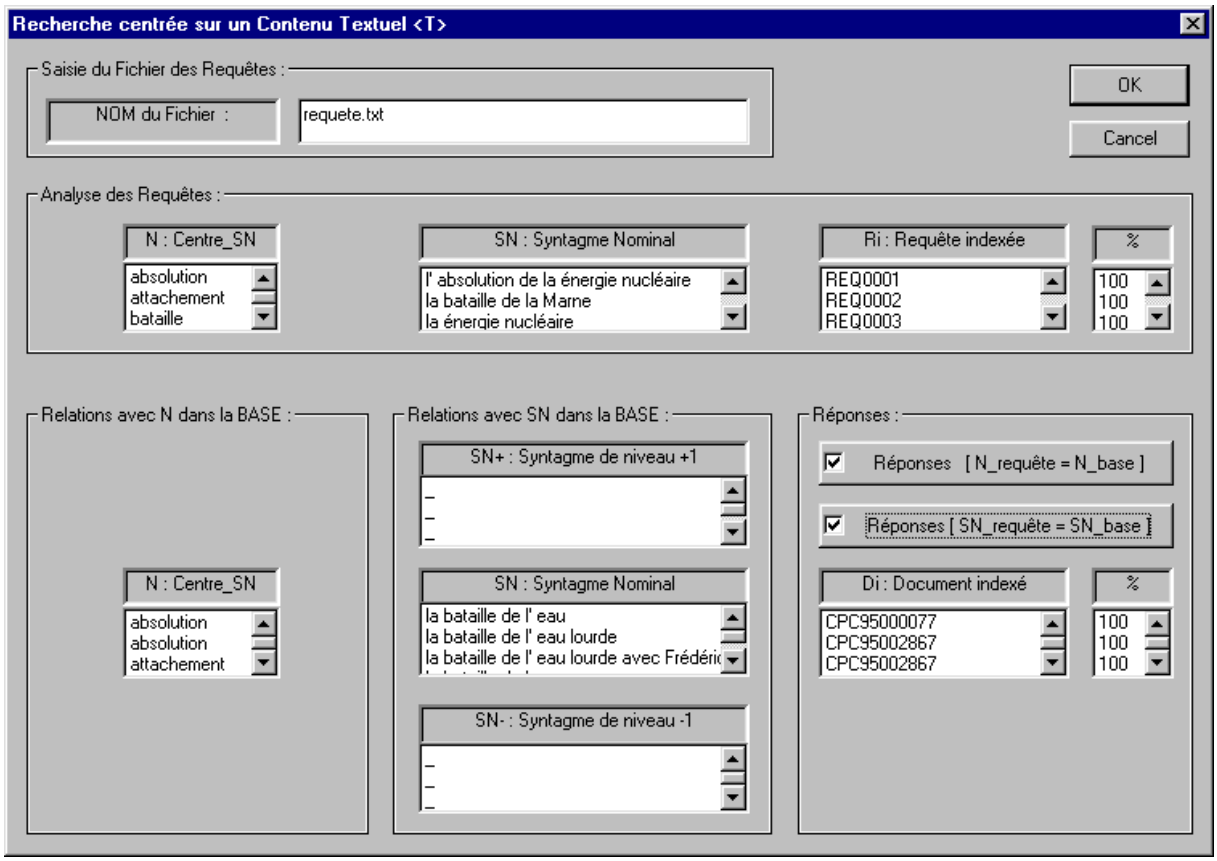
Il s'agit de construire une représentation textuelle (ensemble de phrases) du thème de la recherche qui sera soumis à l'analyse automatique. L'analyseur morpho-syntaxique procède à l'analyse de la requête de manière identique à celle d'un résumé sur un document.

- Analyse automatique du descriptif textuel : Requêtes d'interrogation



A la fin de l'analyse morpho-syntaxique du texte d'interrogation, la nouvelle représentation de la requête contient des connaissances autour du syntagme nominal. Cette représentation va permettre de retrouver les centres et les syntagmes nominaux équivalents dans le fichier inverse de la base et ainsi de donner les références aux documents concernés.

- Recherche dans la base : liaison entre requête et fichier inverse

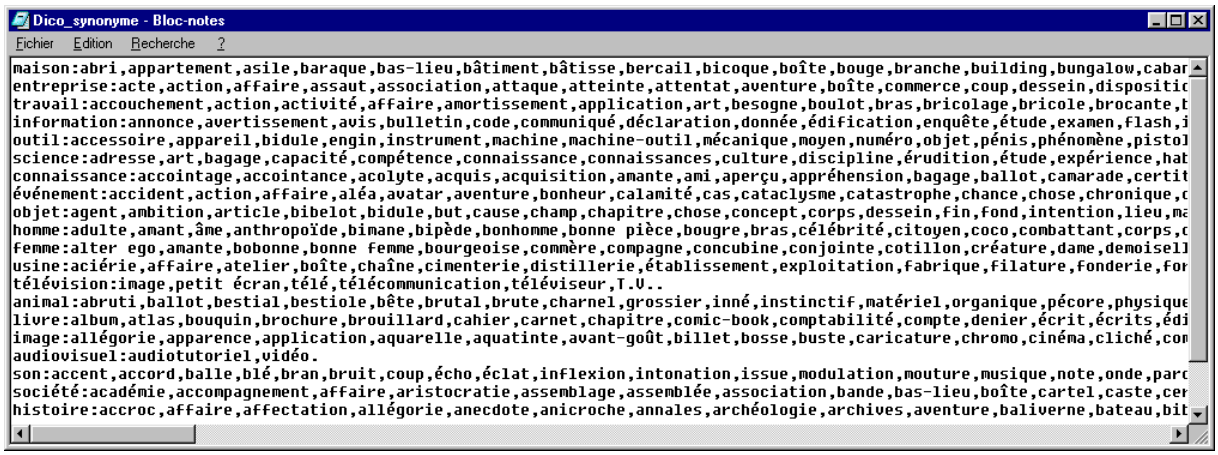


L'interface graphique de la recherche d'information centrée une requête textuelle consiste d'abord à donner la référence du fichier d'analyse de la requête [**NOM du Fichier**], puis d'appliquer la recherche par [OK]. Les connaissances autour des SN-requête sont affichées dans le groupement [**Analyse des Requêtes**] . A la suite de cette représentation, une comparaison automatique est opérée entre la requête et la base. Il s'agit de vérifier si N_requête = N_base et si SN_requête = SN_base.
Les réponses données à la fin de cette comparaison consiste à donner les références des documents dans la base qui vérifient ces conditions.

A.3.6- Recherche basée sur les synonymes du centre nominal

Pour remédier aux échecs de la recherche d'information dans la base, une solution consiste à opérer la recherche sur les synonymes (N_{syn}) d'un centre nominal N.

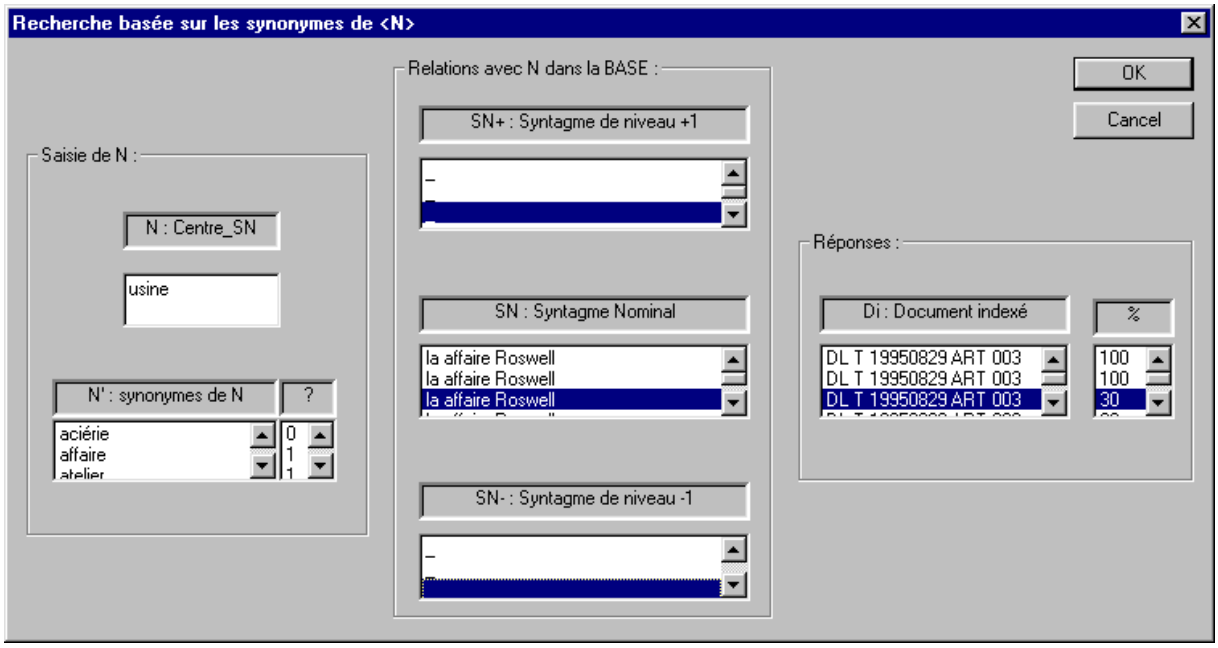
- Emploi d'un dictionnaire de synonymes : dictionnaire échantillon



Un dictionnaire échantillon des synonymes est constitué pour cet effet. Il consiste à donner pour chaque centre nominal (prédicat libre) N une liste de synonymes.
La structuration de cette base de synonymes est $[N_i]$: $[N_{syn\_i1}]$, $[N_{syn\_i2}]$, …, $[N_{syn\_in}]$

Exemples : zone : région, secteur, lieu, espace.
mère : cause, origine, source, maman ; etc.

- Recherche par les synonymes dans la base d'information :



L'interface graphique de la recherche par les synonymes consiste dans son mode opératoire à saisir le centre nominal N puis d'appliquer la recherche par [OK]. Un premier traitement consiste à consulter le dictionnaire des synonymes et de retrouver les synonymes N_{syn} de N.

Le second traitement consiste à opérer successivement sur la liste des synonymes représentée pour retrouver dans la base les documents associés aux synonymes.

A.4- Conclusion

L'analyse et le développement de tels outils n'étaient pas faciles. Il était nécessaire de leur assurer l'aspect modulaire et de généricité.
L'aspect modulaire réside dans une conception qui programme les échanges entre les objets communicants. Quant à la généricité, celle-ci repose sur une technique de développement permettant de rajouter, de modifier ou d'intégrer facilement de « nouveaux comportements » aux objets.

Notre exploration de ces techniques a commencé avec l'étude du formalisme ATN. Certains aspects de ce formalisme ont été analysés dans le but de pouvoir généraliser les protocoles du modèle. Ces protocoles ont été employés afin de définir le mécanisme des objets dans le modèle.
De manière générale, le modèle ATN a la puissance d'une machine de Turing (et non décidable pour l'arrêt), mais certaines restrictions apportées au modèle font que le formalisme devient décidable (*Tab.1*).

Modèle	Système/LSPL *(langage spécialisé)*	LI *(implémentation)*	Caractéristiques	Références
ATN	ATN	LISP	Extension des automates à pile	Woods, 1970
	CATN	LISP	Cascade d'ATN	Woods, ~1980
	REZO *(ATN décidable)*	Pascal	Spécialisation et généralisation des ATN	Stewart, 1978
	AL/DLT	C	Extension des ATN	Doedens et Zuijlen, 1988

Tab.1: Quelques représentations du formalisme ATN.

Dans un cadre plus général sur l'implémentation, il s'agit d'une extension dans le « *génie logiciel* » pour le « *génie linguiciel* ». Dans cette extension, il ne suffit pas d'offrir des « boîtes à outils », mais aussi de construire des « familles d'outils », c'est-à-dire stratifier les différents niveaux d'une application, définir les niveaux d'utilisation et construire selon une approche modulaire les objets à manipuler tout en programmant les protocoles d'échange entre eux.
Selon cette approche d'implémentation, nous avons pu construire nos outils et développer notre Plate-forme d'analyse automatique du langage et de recherche d'information. L'ensemble de ces outils automatiques et les extensions adoptés dans les modules implémentés convergent vers la gestion et la manipulation des connaissances issues des informations écrites.

BIBLIOGRAPHIE

[ABEILLÉ, 95] Anne Abeillé. Syntaxe et TALN. *Ecole d'été CNET (5, Trégastel 1995) – Traitement des langues naturelles*, Editions CNET, pp. III.1-11.

[ABEILLÉ, 98] Abeillé, A. .Grammaires génératives et grammaires d'unification. *Langages*, 1998, 129p.

[AGIRRE, 96] E. AGIRRE, X. ARREGI, X. ARTOLA, A. DIAZ DE ILLAROZA, K. SARASOLA, A. SOROA. Constructing an intelligent dictionnary help system. *in Natural Language Engineering,* 1996, vol.2, n°3, Cambrige university press, pp.229-252.

[AHLSWEDE, 88] Ahlswede T., et *al*. .Automatic construction of a phrasal thesaurus for an information retrieval system from a readable dictionary. *RIAO'88 Recherche d'Information Assistée par Ordinateur*. MIT Combridge, March 1988, p.597-608.

[ALLOTT, 2001] Robin Allott. The Structural Inter-relation of Language, Visual Perception and Action - THE NATURAL ORIGIN OF LANGUAGE. ABLE Publishing : Herts. UK, April 2001.

[On-line]- <URL : http://www.percepp.demon.co.uk/alphabet.htm >

[ALLOUCHE, 99] J. Allouche, X. Casanova, R. Estivals, M. Porada, F. Richaudeau, R. Risler, Introduction à la schématique. *Revue de Bibliologie : Schéma et Schématisation*, 1999, n°45, p.49-61.

[AMAR, 92] AMAR Muriel. *Une étude de cas : le progiciel Termino*. Mémoire de DESS en Information et Documentation, Paris, Institut d'Etudes politiques de Paris/CSSID, septembre 1992.

[AMAR, 97] AMAR Muriel. *Les fondements théoriques de l'indexation : une approche linguistique*. Thèse de Doctorat en Science de l'Information et de la Communication : Université Lumière Lyon 2. 1997, 410p.

[AMSALI, 98] Amsali P., Bras M. .DRT et Compositionnalité. *in TAL Journal,* 1998, vol.39, n°1, p.131-160.
[On-line]- <URL : http://talana.linguist.jussieu.fr/~amsili/Rech/Publis.html >

[ANDREEWSKY, 75] Andreewsky A., Combrisson F., Fluhr C. .*Le problème de l'identification automatique de concepts*. Rapport d'études : Note CEA-N-1816, 1975.

[ANDREEWSKY, 76] Andreewsky A., Fluhr C. .*Indexation automatique - maintenance et gestion d'un système documentaire - 1ère partie: aspects théoriques*. Rapport d'études : Note CEA-N-1694(1), 1976.

[ANDREEWSKY, 77] Andreewsky A., Fluhr C., Debili F. .Computational learning of semantic lexical relations for the generation and automatical analysis of content. *IFIP Congress Proceedings*, 1977, p.667-672.

[ANDREEWSKY, 96] Alexandre Andreewsky. Les systèmes documentaires SPIRIT et MICRO-MIND. *Informatique Documentaire* : *Bulletin du centre de hautes études internationales d'informatique documentaire*, Mars 1996, N°61, Ed. CID Paris, p.29-41.

[ANDRIEU, 96a] Andrieu Olivier. INTRANET: la révolution interne. *in Technologies Internationales*, Mai 1996, N°. 24, p. 31-34.

[ANDRIEU, 96b] Andrieu Olivier. Informatique et réseaux : Java mène la danse. *in Technologies Internationales*, octobre 1996, N°.28, p.33-36.

[ANTON, 88] JP.ANTON, F.DAGORRET, F.LARRIEU. *The AQUEDUCT system.* in RIAO'88 Recherche d'Information Assistée par Ordinateur, p.51-64, MIT, Cambridge,

Massachussetts, USA, Mars 1988.

[BACHIMONT, 2000] Bachimont Bruno. Indexation audiovisuelle : une problématique en pleine évolution. *L'OBJET : logiciel, base de données, réseaux* **:** Volume 6, 2000
[On-line]- <URL : http://sunsite.informatik.rwth-aachen.de/dblp/db/journals/Lobjet/ >

[BACHIMONT, 92] Bachimont B. *.Le contrôle dans les systèmes à base de connaissances : contribution à l'épistémologie de l'IA*. Paris : Hermès, 1992, 317p.

[BACHIMONT, 99a] Bachimont Bruno. La documentation au coeur du processus de production. *in Dossier de l'Audiovisuel*, Janvier-Février 1999, n°83, INA-Publications, p.38-39.

[BACHIMONT, 99b] Bachimont Bruno. Des problématiques ouvertes – la complexité audiovisuelle : enjeux pour une recherche interdisciplinaire. *in Dossier de l'Audiovisuel*, Mai-Juin 1999, n°85, INA-Publications, p.49-51.

[BACHIMONT, 99c] Bachimont Bruno. Bibliothèques numériques audiovisuelles : des enjeux scientifiques et techniques. *Document Numérique*, 1999, Vol.2, n° 3-4, p.219-242.

[BALPE, 95] Balpe Jean-Pierre, Lelu Alain, Saleh Imad. *Hypertextes et hypermédias : Réalisations, outils et méthodes*. Paris : HERMÈS, 1995, 317p.

[BARRAUD, 96] Barraud Philippe. Choix de grammaire et organisation du lexique. *Lexicomatique et Dictionnairiques (actes Vème Journées scientifiques du réseau thématique Lexicologie, terminologie et traduction) :* Lyon, Septembre 1996, pp.379-395.

[BASSANO, 81] Jean-Claude Bassano. Des systèmes d'information de type documentaire et factuel. *in Textes des Communications IDT'81*, 1981, Ed. adbs & anrt Paris, pp.245-249.

[BATES, 78] Madeleine Bates. The Theory and Practice of Augmented Transition Network Grammars. in *Natural Language Communication with Computers 1978*, p.191-259.

[On-line]- <URL : http://ftp.informatik.rwth-aachen.de/dblp/db/conf/db-workshops/nl78.html >

[BATES, 83] Madeleine Bates, Robert J. Bobrow. Information Retrieval Using a Transportable Natural Language Interface. *in Proceedings SIGIR 1983*, p.81-86.

[On-line]- <URL : http://ftp.informatik.rwth-aachen.de/dblp/db/conf/sigir/sigir83.html >

[BAXTER, 96] Graeme Baxter, Douglas Anderson. Image indexing and retrieval : some problems and proposed solutions. *in Internet research : electronic networking applications and policy.,* 1996, Vol. 6, n°4, p.67-76.

[BELLOT, 97] P. Bellot, M. El-Bèze. Un algorithme de segmentation automatique de corpus : le système SIAC. *Actes JST Francil 1997*, Editions AUPELF-UREF, (Avignon) France, 1997, p.113-117.

[BELTRAMETTI, 95a] Monica Beltrametti, Laurent Julliard, Françoise Renzetti. *Recherche d'information et bibliothèque virtuelle: le modèle de Callimarque*. Publications Rank Xerox Research Center, Meylan France, 1995, 10p.

[BELTRAMETTI, 95b] Beltrametti M., Julliard L., Renzetti F. .Recherche d'information et bibliothèque virtuelle, le modèle de Callimaque. *Le Micro-bulletin*, n°59, mai/juin 1995.

[BEN ABDALLAH, 95] Nabil Ben Abdallah. Description de documents textuels : indices pour une typologie prenant en compte le contexte et la finalité de la communication. *Les cahiers des sciences de l'information et de la communication,* Février 1995, n°4, Publications Universités Grenoble 2 et Lyon 2, p.3-22.

[BENABOUD, 92] A. Benaboud, M. Billaud. *Production de documents structurés sous forme SGML*. Rapport de Synthèse PFE : INSA-Lyon et EDF-GDF Clamart,1992, 5p.

[BENARD, 95] Philippe Bénard, Alain Dang-Van-Mien. *L'architecture des services : Pratiquer et maîtriser les systèmes ouverts*. France : Addison-Wesley Editions, 1995, 200p.

[BENMASSAOUD, 91] Mohammed Benmassaoud. *Contribution à l'analyse automatique des analogies dans le lexique - Programmation en Prolog et en Bases de données*. Mémoire de DEA : Université Lyon 2, 1991, p.6-27.

[BENVENISTE, 74] BENVENISTE Émile. *Problèmes de linguistique générale 2.* Paris : Collection Tel., Editions Gallimard, 1974, 286 p.

[BERRENDONNER, 90] BERRENDONNER Alain, *Grammaire pour un analyseur : aspects morphologiques*. Les Cahiers du Criss. Grenoble : Centre de Recherche en Informatique appliquée aux Sciences Sociales. Université des Sciences Sociales de Grenoble, Nov. 1990, 88 p.

[BERRUT, 86] Catherine Berrut, Patrick Palmer: Solving Grammatical Ambiguities within a Surface Syntactical Parser for Automatic Indexing. *in Proceedings* SIGIR 1986: 1986, p.123-130.

[BERRUT, 89] Catherine Berrut, Yves Chiaramella. Indexing Medical Reports in a Multimedia Environment: The RIME Experimental Approach. *in Proceedings SIGIR 1989:* **1989,** *p.187-197.*

[BERTHOUZOZ, 96] Cathy Berthouzoz. *Modélisation orientée-objets d'une grammaire de type - Gouvernement et Liage – dans le cadre d'un système de traduction multilingue*. Notes techniques 1996/n°5, Laboratoire d'analyse et de technologie du langage : université de Genève, 1996, 60p.

[BERTHOUZOZ, 97] C. Berthouzoz, P. Merlo. Filtrage structurel versus filtrage statistique : expériences sur la résolution de l'ambiguïté syntaxique. *in Actes JST Francil* 1997, Editions AUPELF-UREF, (Avignon) France, p.353-359.

[BERTRAND, 2001] Patrice BERTRAND, Edwin DIDAY, Yves LECHEVALLIER, Marie CHEVANT, Ghazi BEL MUFTI, Edgard MFOUMOUNE. Projet CLOREC : Classification automatique – ordres conjugués et classification Pyramidale [On-line]. [01.06.01] Available from internet :

[On-line]- <URL : http://www.inria.fr >

[BESSER, 95] Howard Besser. Getting the picture on image databases : the Basics. *in DATABASE*, April/May 1995, p.12-19

[BICHARD, 92] J-P. Bichard. *Les réseaux d'entreprise : accéder, partager, échanger*. Paris : DUNOD, 1992.165p.

[BIÉBOW, 91] B. Biébow, S. Szulman. Interprétation de groupes nominaux complexes dans un réseau sémantique KLONE. *8eme Congrès AFCET – reconnaissances des formes et IA*, Lyon novembre 1991, vol.2, Editions AFCET, p.887-892.

[BIÉBOW, 97] B. Biébow, S. Szulman. Avancée sur le concept de base de connaissances terminologique. *in PRC-GDR IA'97 Actes des 6eme journées nationales* , Ed. HERMÈS, Paris : 1997, p.357-370

[BLANQUET, 93] Marie-France Blanquet. La fonction documentaire : Etude dans une perspective historique. *Documentaliste et Sciences de l'information*, vol.30/n°4-5, 1993, p.199-204

[BLANQUET, 94] Marie-France Blanquet, *Intelligence artificielle et systèmes d'information : la langage naturel*, Paris : ESF éditeur, 1994.

[BONFANT alii, 94] L. Bonfante et alii. *La naissance des écritures : du Cunéiforme à l'alphabet*. Edition SEUIL, Paris 1994, 503.p.

[BOOKSTEIN, 75] Bookstein A., Swanson D.R. .A decision theoritic foundation of indexing. *Journal of the American Society for Information Science*, 1975, vol. 26, p.45-50.

[BORIAS, 95] Philippe Borias. *Système de gestion de la documentation sur le WWW*. Projet de PFE, INSA-Lyon, 1995, 5p.

[BOUCHART, 95] L.H. Bouchart, L. Emirkanian. Elaboration d'un dictionnaire informatisé pour le traitement automatique de la langue. *Lexicomatique et Dictionnairiques (actes IVème Journées scientifiques du réseau thématique Lexicologie, terminologie et traduction)*, Lyon, Septembre 1995, pp.59-76.

[BOUCHÉ, 88] Bouché R. .Valeur référentielle et langage d'indexation. *Colloque Archives et temps réel, CEDRO-Université Lille III*, Novembre 1988, Ed. ADBS Nord.

[BOUCHÉ, 89] Richard Bouché. Le syntagme Nominal - une nouvelle approche des bases de données textuelles. *in META,* 1989, vol. XXXIV, N°3, pp.428-434.

[BOUCHÉ, 90] BOUCHÉ R., LAINÉ S., METZGER J-P. .Extraction des connaissances à partir d'une collection de documents. in : *Tools of knowledge organization and the human interface*, Congrès organisé par l'ISKO (International Society for Knowledge Organization), Darmstadt (D), 14-17 Août 1990.

[BOUCHÉ, 91] Richard Bouché, Nicolas Germain. Bibliométrie, infométrie et analyse automatique de documents écrits. *in Revue française de bibliométrie,* 1991, n°9, p.352-366.

[BOUGHANEM, 92] Boughanem M., Soule-Dupuy C. .Un modèle connexionniste pour la recherche d'informations. in *L'informatique Documentaire :* Bulletin du centre de hautes études internationales d'informatique documentaire, Sep.1992, N°.47*,* p.13-30.

[BOURNAUD, 2000a] Bournaud, I. and Courtine, M. and Zucker J.-D. .Abstractions for knowledge Organization of Relational Descriptions. *In LNAI* : *Symposium on Abstraction Reformulation and Approximation SARA 2000*, July 2000.
[On-line]- <URL : http://www.lri.fr/ia/isabel/biblio-date.fr.html >

[BOURNAUD, 2000b] Bournaud I. and Courtine M. **.**KIDS : Un algorithme itératif pour l'organisation de données relationnelles. in *Conférence d'Apprentissage CAp'2000*, June 2000.
[On-line]- <URL : http://www.lri.fr/ia/isabel/biblio-date.fr.html >

[BOURNAUD, 2000c] Bournaud I. .Automatic Objects Organization. in *Workshop of ECOOP'2000, Objects and Classification : a Natural Convergence*, June 2000., p. 47-56
[On-line]- <URL : http://www.lri.fr/ia/isabel/biblio-date.fr.html >

[BOURNAUD, 96] Bournaud I. *.Regroupement conceptuel pour l'organisation de connaissances.* Thèse de Doctorat, Spécialité Informatique : Université Paris VI, 1996.
[On-line]- <URL : http://www.lri.fr/ia/isabel/biblio-date.fr.html >

[BOURNAUD, 97] I. Bournaud. Construction itérative et classifications conceptuelles. in *5eme Rencontres de la société Francophone de la Classification.* Sep. 1997, Lyon-France.

[BOURNAUD, 98a] Bournaud, I. and Mathieu J. .Le regroupement conceptuel pour l'élicitation de connaissances utilisées dans la construction de classifications. In N. Bacri *alii*, editor, *Catégorisation*, Collection Psychologie de la pensée. P.U.F. : Paris, 1998.
[On-line]- <URL : http://www.lri.fr/ia/isabel/biblio-date.fr.html >

[BOURNAUD, 98b] I. Bournaud, J-D. Zucker. Prendre en compte la structure des descriptions dans la construction de classifications conceptuelles. in *6eme Rencontres de la Société Francophone de Classification,* Septembre1998, Montpellier France

[BOURSIER, 94] Patrice Boursier, Pierre-Antoine Taufour. La technologie multimédia. Paris : Edition HERMÈS (2nd Edition), 1994, 249p.

[BOUSQUET, 91] Réné Bousquet. La recherche documentaire et l'image. in *Textes des Communications IDT'91*, 1991 : Bordeaux, Ed. adbs & anrt Paris, p.227-228

[BOUTIÉ, 96] Philippe Boutié. Ceci tuera-t-il cela ? Les média numériques vont-ils révolutionner la communication ? . in *Revue Française du marketing*, 1/1996, N°.156, p.5-10

[BRILL, 92] Brill F.R., Brown D.E., Martin N. .Fast Genetic Selection of Features Neural Network Classifiers. *IEEE Transaction on Neural Networks*, Mars 1992, Vol.3, N°2.

[BROWNE, 96] Glenda Browne. Automatic indexing and abstracting. *in Indexing in Electronic Age Conference*, Robertson, NSW 20-21 April 1996, Australian Society of Indexers, 8p.

[BRUN, 2000] *Caroline Brun.* Coordination et analyse automatique du français écrit dans le cadre de la Grammaire Lexicale Fonctionnelle. Publications internes Xerox, 2000
[On-line]- <URL : http://www.parc.xerox.com/istl/groups/nltt/pargram/ >

[CARMONA, 98] J. Carmona, S. Cervell, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé & J. Turmo. An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. in *First International Conference on Language Resources and Evaluation (LREC'98)*. Granada : Spain, 1998.
<URL : http://www.lsi.upc.es/~nlp/pap.html >

[CASANOVA, 99] Xavier Casanova. Vers la schématologie. *Revue de Bibliologie : Schéma et Schématisation*, 1999, n°45, p.39-47.

[CAVAZZA, 92] Marc Cavazza, Pierre Zweigenbaum. Compréhension automatique du langage naturel par construction de modèles. *Technique et Sscience Informatiques*. Vol.11 – n°4/1992, p.119-138.

[CHAMPENIER, 96] Thiébaut Champenier, David Pautet. *Mise à disposition à travers le réseau Internet de la littérature grise produite à l'INSA*. Projet de PFE, 1996 , INSA-Lyon, 65p.

[CHARDENON, 97] C. Chardenon. Outil d'étiquetage parenthésé : Applications à la recherche d'informations. *Note Technique CNET-France Télécom : NT/DSM/GR/01, 1997, 40p.*

[CHAUDIRON, 94] Stéphane Chaudiron. L'intégration des technologies d'ingénierie linguistique dans le traitement de l'information. in *Textes des Communications IDT'94*, 1994, Paris, Ed. adbs & anrt Paris, p.100-104.

[CHAUMIER, 90] Jacques Chaumier, Martine Dejean. L'indexation documentaire – de l'analyse conceptuelle humaine à l'analyse automatique morphosyntaxique. *Documentaliste.* novembre-décembre 1990, Vol.27, n°6, p. 275-279.

[CHAUMIER, 92] J. Chaumier, M. Déjean. L'indexation assistée par ordinateur : principes et méthodes.

in *Documentaliste - Sciences de l'information*, 1992, Vol.29/n°1

[CHAWK, 93] CHAWK Mohamad. *La réécriture de D' - les déterminants complexes du français : lexique et syntaxe.* Mémoire de DEA, enssib - Lyon, 1993, 108p.

[CHEN, 98a] H. Chen, J. Martinez, A. Kirchhoff, T. D. Ng, B.R. Schatz..Alleviating Search Uncertainly through Concept Associations : automatic indexing, co-occurrence analysis, and parrallel computing. in *Journal of the American Society for Information Science.* 1998, vol. 49(3) : p.206-216.

[CHEN, 98b] H. Chen, Y. Zhang, A.L. Houston. Semantic indexing and searching using a Hopfield net. in *Journal of Information Science.* 1998, vol.24 (1), pp.3-18.

[CHIARAMELLA, 86] Yves Chiaramella, Bruno Defude, Marie-France Bruandet, D. Kerkouba: IOTA: A Full Text Information Retrieval System. SIGIR 1986: 1986, p.207-213

[CHOMSKY, 86] *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger Publishers, 1986.

On-line]- <URL : http://web.mit.edu/linguistics/www/chomsky.home.html >

[CHOMSKY, 87] *Language and Problems of Knowledge: The Managua Lectures*, Cambridge: MIT Press, 1987.

On-line]- <URL : http://web.mit.edu/linguistics/www/chomsky.home.html >

[CHUNG, 98] Yi-Ming Chung, William M. Pottenger, Bruce R. Schatz. *Automatic Subject Indexing Using an Associative Neural Network.* CANIS - Community Systems Lab University of Illinois at Urbana-Champaign, Champaign, IL 61820

[On-line]- <URL : http://www.canis.uiuc.edu/projects/interspace/technical/canis-report-0003/index.html > [last updated 05-09-98]

[CLAVEL, 93] G. Clavel, F. Walther, J. Walther. Indexation automatique de fonds bibliothéconomiques. *in ARBIDO-R,* Vol.8 (1993) n°1, Suisse, p.14-19.

[CROFT, 91] W. Bruce Croft, Howard R. Turtle, David D. Lewis. The Use of Phrases and Structured Queries. in *Information Retrieval. SIGIR 1991: p.32-45.*
[On-line]-
<URL : >

[CULTURE, 82] Naissance de l'écriture : Cunéiformes et Hiéroglyphes. Ministère de la Culture de R.France, Edition des musées nationaux, 1982, 381.p.

[CUNEIFORME, 2001] [On-line] *Ecriture cunéiforme.* BnF, France, 2001

http://www.bnf.fr/web-bnf/pedagos/dossiecr/sp-cune3.htm

[DACHLET, 90] Roland Dachlet. Etat de l'Art de la recherche en informatique documentaire : la représentation des documents et l'accès à l'information. *Rapport de recherche de l'INRIA- Rocquencourt, Avril 1990, Projet : PSYCHO-ERGO - 32 p.*
[On-line]- < URL : http://www.inria.fr/rrrt/rr-1201.html >

[DAGOGNET, 73] François Dagognet. *Ecriture et Iconographie*. Edition Lib. Philosophique J.VRIN, Paris 1973, 170.p.

[DAVALO, 92] Davalo E., Naïm P. .*Des réseaux de neurones*. Paris : Eyrolles, 1992, 232p.

[DE BRITO, 91] De Brito M. .*Réalisation d'un analyseur morpho-syntaxique pour la reconnaissance du syntagme nominal : Utilisation des grammaires Affixes*. Thèse de Doctorat : Université Lyon 1, 1991, 220p.

[DE BRITO, 92] DE BRITO Marcilio. Information System in natural languages: the search for an automatic indexing system. in *Ciência da Informação*. 1992, vol.21, n°.3. p.223-232.

[DÉBILI, 82] Débili F. .*Analyse syntaxico-sémantique fondée sur une acquisition automatique de relations lexicales-sémantiques*. Thèse de Doctorat d'Etat : Université Paris XI, 1982.

[DEGEZ, 70] Danièle Degez. Conclusions de l'étude expérimentale des méthodes d'analyse et d'indexation SEMA-ORTF. *Publications Archives audiovisuelles de la télévision*, Novembre 1970.

[DENOUE, 2000] Laurent Denoue, Laurence Vignolet. L'importance des annotations : Application à la classification des documents du web. in *Document Numérique*, 2000, Ed. Hermès, Vol.4, N°.1-2, p.37-57

[DESAI, 95] Bapin C. Desai. Indexing and Searching Virtual Libraries. in *CIC Forum: America in the Age of Information*, July 1995, Bethesda MD
[On-line]- <URL : http://www.cs.concordia.ca/~faculty/bcdesai/ >

[DEWEZE, 81] DEWEZE André. .*Réseaux sémantiques : essai de modélisation, application à l'indexation et à la recherche de l'information documentaire*. Thèse de Doctorat : Informatique Documentaire : Université Claude Bernard Lyon 1, 1981.

[DEWEZE, 93] DEWEZE A. *Informatique documentaire*. 4e éd. Paris : Masson, 1993, 292 p.

[DHENIN, 96] Cyril Dhénin. Internet : Les robots déchiffrent le WEB. in *le monde informatique*, date : 22 Mars 1996, p.24-25.

[DIDAY, 2000] Diday, E. and Kodratoff, Y. and Brito, T. and Moulet M. .Induction symbolique-numérique à partir de données. Paris : Editions Cepadues, 2000.
[On-line]- <URL : http://www.lri.fr/ia/yk/biblio-date.fr.html >

[DIDAY, 2001] Edwin Diday, Patrice Bertrand, Edgard Mfoumoune. .Classification Pyramidale : Une nouvelle implémentation de la CAP.
[On-line]- <URL : http://www.inria.fr/rapportsactivite/RA95/clorec/node14.html >

[DIENG, 94] Rose Dieng. *Méthodes et outils d'acquisition des connaissances.* Rapport d'activité 1994, Projet ACACIA -
[On-line]- <URL : http://www.inria.fr/rapportsactivite/RA94/ACACIA.html >

[DIXON, 97] Mark Dixon. *An overview of Document Mining Technologie.* October 4, 1997.
[On-line]- <URL : http://citeseer.nj.nec.com/dixon97overview.html >

[DJEBAR, 89] Tahar Djebar, Eliane Bonfils, Sylvia Seidel, Gilles Taladoire. .Des systèmes documentaires relationnels au concept d'hyperbase. in *Textes des Communications IDT'89*, 1989, Ed. adbs & anrt Paris, p.221-227.

[DOYLE, 62] Doyle L.B. .Indexing and abstracting by association. *American Documentation,* Oct. 1962, Vol.13, n°4, p.378-390.

[DUPONT, 83] Dupont P. .*Eléments logico-sémantiques pour une analyse du français*. Thèse de Doctorat en linguistique : Université Lyon 2, 1983.

[EGYPTE, 98] *Civilisation égyptienne – Ecriture.* Musée canadien des civilisations.
[On-line]- <URL : http://www.civilization.ca/members/civiliz/egypt >

[EGYPTE, 99] *Ecriture égyptienne : une image du monde.* BnF France, Janvier 1999,

[On-line]- <URL : http://www.bnf.fr/web-bnf/pedagos/dossiecr/in-hiero.htm >

[EL GUEDJ, 97] El Guedj, P. Nugues. *Analyse syntaxique combinant deux formalismes au sein d'un Chart Hiérarchique P-O.*. Publication interne Laboratoire GREYC, Univ. Caen, 1997. [On-line]- <URL:http://www.info.unicaen.fr/ñugues/Articles/grkg1997/grkg1997.html>

[ELGHOUL, 98] M. Elghoul. Processus d'indexation et de recherche d'information dans les systèmes multimédia. in *Informatique Documentaire,* n°71, Décembre 1998, p.29-43.

[FALCIGNO, 95] Kathleen Falcigno, Tim Green. Home page-Sweet : creating a Web Presence, in *Database*, April/May 1995, p.20-28.

[FARAJ, 96] Najib Faraj, Robert Godin, Rokia Missaoui. Analyse d'une méthode d'indexation automatique basée sur une analyse syntaxique du texte. in *Revue Canadienne des Sciences de l'Information et de bibliothéconomie,* April 1996, Vol. 21, N°. 1, 21p.

[FARWELL, 93] David Farwell, Loise Guthrie, Yorick Wilks. Autmatically creating lexical entries for ULTRA a multilingual MT system. in *MT*, 1993, N°8, Kluwer Academic Publishers, p.127-145.

[FAURE, 98] D. Faure and C. Nédellec. Apprentissage de cadres de sous-catégorisation et de restrictions de sélection à partir de textes. Pierre Zweigenbaum : editor, in *5ème conférence annuelle sur le traitement automatique des langues naturelles (TALN'98)*, , Juin 1998, p. 233-235. [On-line]- <URL : http://www.lri.fr/ia/cn/biblio-date.fr.html >

[FAY-VARNIER, 91] C. Fay-Varnier, C. Fouqueré, G. Prigent, P. Zweingenbaum. Modules syntaxiques des systèmes d'analyse du français. in *TSI – Techniques et Science Informatiques*, 1991, vol.10, n°6, Ed. AFCET-Bordas, p.403-425.

[FELDMAN, 96] Susan Feldman. Comparing DIALOG, TARGET and DR-LINK (testing natural language). in *ONLINE,* november-december 1996, vol.71, p.71-79.

[FELDMAN, 98a] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Scheler, O. Zamir. Text Mining at the Term Level. in *Proceedings. of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98)* , Nantes (France), sep. 1998 [On-line]- <URL : http://liawww.epfl.ch/ >

[FELDMAN, 98b] R. Feldman, M. Fresko, H. Hirsh, Y. Aumann, O. Liphstat, Y. Schler, M. Rajman. .Knowledge Management : A Text Mining Approach. *Proceedings of the 2nd Conference on Practical Aspects of Knowledge Management (PAKM98),* Basel - Switzerland, 29-30 October 1998, 9p.

[FERRET, 2000] O. Ferret, B. Grau, G. Illouz, C. Jacquemin. QALC : the Question-Answering program of the Language and Cognition group at LIMSI-CNRS. in *Proceedings 9th Text Retrieval Conference (TREC8)*, 2000, NIST, Gaithersburgh MD.

[FGAIER, 92] Nabil Fgaier. *Des nouvelles méthodes de comparaison question-document pour des systèmes gérant des documents structurés - multimédia et organisés selon une approche orientée objet.* Thèse de Doctorat : Université de Paris-Sud, 1992, p.109-137.

[FLEXER, 91] Annie Flexer. TAURUS: Système de gestion électronique de documents multimédia. in *Textes des Communications IDT'91*, 1991 : Bordeaux, Ed. adbs & anrt Paris, p.225-226.

[FLUHR, 92] FLUHR Christian. *Le traitement du langage naturel dans la recherche d'information documentaire.* Cours INRIA - Interfaces Intelligentes dans l'Information Scientifique et Technique, 18-22 Mai 1992, p.103-128.

[FLUHR, 97] Christian Fluhr. *Techniques d'indexation et applications.* Techniques de base pour le multimédia, Edited by G. Weidenfeld *alii*, MASSON, Paris, 1997, p.65-74.

[FREGE, 71] Frege G. (1892/1971), « *Sens et dénotation* », in Ecrits logiques et philosophiques, Paris, Le Seuil, p.102-126.

[FRIEDRICH, 96] Michel Friedrich. *Aide à la navigation dans les hypertextes par la programmation logique.* Rapport de synthèse du PFE, INSA-Lyon, 1996, 10p.

[FUCHS, 93] Fuchs C. et *al.* .*Linguistique et traitements automatiques des langues*. Paris : Hachette Supérieur, 1993, 304p.

[FUGMANN, 83] Robert Fugmann.*The analytico-synthetic foundation for large indexing & information retrieval systems.* Bangalore : Sarada Ranganathan Endowment for Library Science, 1983, 58 p.

[FUGMANN, 93] R. Fugmann. *Subject Analysis and Indexing: Theoretical Foundation and Practical Advice*. INDEKS VERLAG, Germany, 1993.

[FUGMANN, 94] *Fugmann, Robert. Galileo and the inverse precision/recall relationship.* Knowledge Organization, *Vol. 21, 1994, No. 3, pp. 153-154.*

[GALA PAVIA, 2000] Nuria Gala Pavia. Hétérogénéité des corpus : vers un parseur robuste reconfigurable et adaptable. in *Conférence TALN 2000*, Lausanne 16-18 octobre 2000.

[GANASCIA, 87] J.G. Ganascia. *AGAPE et CHARADE : deux techniques d'apprentissage à la construction de bases de connaissances.* Thèse d'Etat : Université Paris XI-Orsay, France, 1987.

[GARDIN, 74] Jean-Claude Gardin. *Les analyses de discours.* Ed. Collection Zethos, 1974

[GASTALDY, 92] BERTRAND-GASTALDY S., PAGOLA G. (1992). L'analyse du contenu textuel en vue de la construction de thésaurus et de l'indexation assistées par ordinateur - applications possibles avec SATO (système d'analyse de textes par ordinateur). *Documentation et bibliothèques*, 38 (2), avril-juin 1992, p.75-89.
[On-line]- <URL : http://www.ling.uqam.ca/sato/publications/bib_fcar.htm >

[GASTALDY, 95] BERTRAND-GASTALDY, S.; GIROUX, L.; LANTEIGNE, D.; DAVID, C. . La modélisation de l'analyse documentaire: à la convergence de la sémiotique, de la psychologie cognitive et de l'intelligence artificielle_. in *Canadian Association for Information Science; Proceedings of the 23rd Annual Conference / Association canadienne de sciences de l'information*, H.A. OLSON et D.B. WARD (ed.). Edmonton: University of Alberta, School of Library and Information Studies, 1995: p.1-11.
[On-line]- <URL : http://www.ualberta.ca/dept/slis/cais/gastaldy.htm >

[GASTALDY, 96] BERTRAND-GASTALDY, S.; PAQUIN, L.-C.; PAGOLA, G.; DAOUST, F. .Le traitement des textes primaires et secondaires pour la conception et le fonctionnement d'un prototype de système expert d'aide à l'analyse des jugements_. in *Traitement automatique du français écrit: développements théoriques et applications*, L. EMIRKANIAN et L.H. BOUCHARD (éd.). Actes du colloque 'Traitement automatique du français écrit', 62e congrès de l'Acfas (16-20 mai 1994). Acfas, 1996: p.241-276.
[On-line]- <URL : http://www.ling.uqam.ca/sato/publications/bib_fcar.htm >

[GAUSSIER, 97] E. Gaussier, G. Grefentette, M. Schulze. Traitement du langage naturel et recherche d'information : quelques expériences sur le français. *Actes JST Francil 1997*, Editions AUPELF-UREF, (Avignon) France, p. 9-14.

[GEDIMAGE, 96] *GEDIMAGE*. LEXIC : interrogation de bases de données textuelles et de banques d'images en langage naturel. *Publications GEDIMAGE, France : Janvier 1996.*

[GERMAIN, 97] Nicolas Germain. Traitement de l'information et analyse textuelle anominale. *Document numérique*, 1997, Volume 1, n°3, p.339-349.

[GIGUET, 98] Emmanuel Giguet. Méthode pour l'analyse automatique des structures formelles sur documents multilingues. *Thèse de Doctorat : Université de Caen, 1998, 220p.*

[GIRARDI, 94] M.R. Girardi, B. Ibrahim. Automatic indexing of software artifacts. in *The Journal of Systems and Software* : Special Issue on Software Reusability, 1994, 13p.

[GOMEZ, 93] P. Gomez, P. Bichon. *Comprendre les réseaux d'entreprise*. Paris : Eyrolles, 1993, 214p.

[GOODY, 94] Jack Goody. *Entre l'oralité et l'écriture*. Traduit de l'anglais par Denise Paulme et révisé par Pascal Ferroli, Presses universitaires de France : PUF, 1994.

[GOODY, 98] Jack Goody. De l'oral à l'écrit. *Propos recueillis par Nicolas Journet, in Sciences*

Humaines, Mai 1998, N°83, p.38-41.

[GREFENSTETTE, 94] G. Grefenstette, P. Tapanainen. What is a word, What is a sentence ? Problems of Tokenization. in *COMPLEX'94*, July 7-10-1994, Budapest

[GROSS, 96] Maurice Gross, Max Silberztein. Outils de traitement linguistique, applications à l'analyse documentaire. *Ecole d'été CNET (5, Trégastel 1995) – Traitement des langues naturelles*, Editions CNET, p. I-1-24.

[GUIHOT, 94] Patrick Guihot, Jean-François Clair. Problèmes liés à la conception de services multimédia intéractifs. *Actes du Séminaire : Ecrit Image Oral et Nouvelles technologies*, 1993-1994, Publications Université Paris 7, p.93-100.

[GUILBAUD, 95] Elisabeth Guilbaud. Le GED et le texte intégral au service de la photo. *Archimag*, n°80, décembre-janvier 1995, p.52-54.

[GUIMIER, 93] Anne-Marie Guimier-Sorbets. Des textes aux images : accès aux informations multimédias par le langage naturel. *Documentaliste – Sciences de l'information*, 1993, vol.30, n°3, p.127-134.

[GUINCHAT, 90] Claire Guinchat, Michel Menon. *Sciences et techniques de l'information et de la documentation*. Publications UNESCO, 2nd Edition 1990.

[GUYOMARD, 95] Marc Guyomard. Référence et dialogue. *Ecole d'été CNET (5, Trégastel 1995) – Traitement des langues naturelles,* Editions CNET, p. IX.2-24

[HAASE, 96] K. Haase. FramerD : Representing knowledge in the large. *IBM systems journal*, 1996, Vol.35, n°3-4, p.381-397.

[HABERT, 91] Benoît Habert. Spécialiser des règles syntaxiques. *8eme Congrès AFCET – reconnaissances des formes et IA*, Lyon novembre 1991, vol.2, Editions AFCET, p.873-878.

[HABERT, 97] B. Habert, M-L. Herviou-Picard, D. Bourigault, R. Quatrin, M. Roumens. Un outil et une méthode pour comparer deux extracteurs de groupes nominaux. *Actes JST Francil 1997*, Editions AUPELF-UREF, 1997 (Avignon) France, p.509-515.

[HAGEGE, 96] Claude Hagège. *L'homme de paroles : contribution linguistique aux sciences humaines*. Edition Fayard, 1996, 316.p.

[HAMOU-LHADJ, 94] Hamou-Lhadj A. .*Modèle connexionniste pour la prise en compte de l'utilisateur dans un système de recherche documentaire*. Mémoire de DEA : ENSSIB et Univ.Lyon1, Oct. 1994, p.7-27.

[HARRIS, 93] Roy, Harris. La sémiologie de l'écriture. CNRS Edition, Paris 1993, 377.p.

[HARTER, 75] Harter S.P. .A probabilistic approach to automatic key word indexing- Part 1: on the distribution of specialty word in technical litterature - Part 2: An algorithm for probabilistic indexing. *Journal of the American Society for Information Science*, , 1975, vol.26, p.197-206, et p.280-289.

[HENNI, 96] Jamal Henni. .Voix et Vidéo sur l'Intranet. in Réseaux et télécoms, pp. 43-48, Octobre 1996.

[HENRY, 97] Patrick Henry. L'indexation automatique d'images. *Technologies Internationales,* Novembre 1997, n°39, p.7-11

[HESS, 92] HESS Michael. An Incrementally Extensible Document Retrieval System Based on Linguistic and Logical Principes. in. : *SIGIR'92. Denmark*, June, 1992, p. 190-197.

[HODGES, 96] J. Hodges, S. Yie, R. Reighart, L. Boggess. An automated system that assists in the generation of document indexes. in *Natural Language Engineering,* vol.2, n°2, Cambrige University Press, 1996, p.137-160.

[HUDRISIER, 81] Henri Hudrisier. Le traitement de l'image -les banques de données iconographiques de presse et encyclopédiques. in *Textes des Communications IDT'81*, 1981, Ed. adbs & anrt Paris, p.385-390

[JACOBSON, 96] N. Jacobson, W. Bender. Color as a determined communication. *IBM systems journal*, 1996, Vol.35, n°3-4, p.526-538.

[JACQUEMIN, 2000] C. Jacquemin, P. Zweigenbaum. Traitement automatique des langues pour l'accès au contenu des documents. in *Le document Multimédia en Sciences du traitement de l'Information,* Ed. CEPADUES EDITIONS, Toulouse, Editors : J. Le Maitre *alii*, 2000, p.71-109.

[JAKOBSON, 96] JAKOBSON R., Essais de linguistique générale, Editions de Minuit, Paris, 1963.

[JODOUIN, 94] Jodouin J-F. .*Les réseaux de neurones -principes et définitions*. Paris : Hermès, 1994, 124p.

[JOLLY, 93] Muriel JOLLY, Jean-Philippe FAYE. *Conception de bornes interactives multimédia*. Mémoire de DEA, INSA de Lyon, 1993.

[JONES, 91] Jones Susan. *Text and Context document processing and storage.* Springer-Verlag London Limited, 1991, 297p.

[JOUVE, 99] O. JOUVE. Les outils d'analyse et de filtrage d'information : l'exemple du projet SAMPLER. *Pôle Universitaire L. De Vinci*, Paris, 1999, p.11
[On-line]- <URL : http://www.idt.fr/pages_fra/actes/actes99/acte35.html >

[KABBAJ, 91] A. Kabbaj. *Intelligence Artificielle en LISP et PROLOG - Chapitre11: Analyse syntaxique.* Paris : Ed. Masson, 1991, p.257-281.

[KAHLAL, 92] Karim Kahlal. *Conduite et évaluation d'une étude préalable relative à un projet d'indexation automatique.* Rapport de Stage DESS Informatique Documentaire, 1992, ENSSIB de Lyon, 54p.

[KALLAS, 87] Kallas G. *Résolution des solutions multiples en analyse morphologique automatique des langues naturelles : modélisation par les chaînes de Markov*. Thèse de doctorat, Université des sciences sociales de Grenoble, 1987.

[KAWKELL, 96] Tony Kawkell, *The multimedia handbook*, USA : Ed. Routledge, 1996.

[KAZMAN, 95] Rick Kazman, William Hunt, Marilyn Mantei. Dynamic Meeting Annotation and Indexing. *Proceedings of the 1995 Pacific Workshop on Distributed Multimedia Systems*, (Honolulu, HI), March 1995, p.11-18.

[KAZMAN, 97] Rick Kazman, John Korminek. *Supporting the Retrieval Process in Multimédia Information Systems.* Departement of computer science : university of Waterloo, 1997

[KIMMEL, 96] Stacey Kimmel. Robot generated databases on the WWW. in *Database,* February-March 1996, p.41-49.

[KODRATOFF, 2000a] Kodratoff Y. .Quelques contraintes symboliques sur le numérique en ECD et en ECT. in *Lecture Note in Computing Science*, 2000.
[On-line]- <URL : http://www.lri.fr/ia/yk/biblio-date.fr.html >

[KODRATOFF, 2000b] Kodratoff Y. .*About Knowledge Discovery in Texts: A definition and an Example*. Technical report Lab. LRI, 2000.
[On-line]- <URL : http://www.lri.fr/ia/yk/biblio-date.fr.html >

[KODRATOFF, 96] Kodratoff Y. .L'extraction de connaissance à partir des données : un nouveau sujet pour la recherche scientifique. in *XIV Congrès INFORSID,* June 1996, *Bordeaux*, p.3-22.
[On-line]- <URL : http://www.lri.fr/ia/yk/biblio-date.fr.html >

[KURAMOTO, 95] KURAMOTO Hélio. *Maquette d'un système de recherche d'information en utilisant des syntagmes nominaux.* Mémoire du DEA. École Nationale Supérieure des Sciences de l'Information et des Bibliothèques, Villeurbanne, 1995.

[KURAMOTO, 99] KURAMOTO Hélio. Proposition d'un Système de Recherche d'Information Assistée par Ordinateur - Avec application à la langue portugaise. *Thèse de doctorat. Lyon : Université Lumière – Lyon 2, 1999. 286 p.*

[LAINÉ, 82] Lainé S. .*Extraction et selection de descripteurs complexes dans un ensemble de textes pour leur indexation automatique*. Thèse de Docteur-Ingénieur en mathématiques (Informatique) : U.C.B. Lyon1, 1982.

[LALLICH, 86] LALLICH-BOIDIN Geneviève. *Analyse syntaxique automatique du français : Applications à l'indexation automatique.* Thèse de doctorat : Université des Sciences

Sociales de Grenoble et Ecole Nationale Superiéure des Mines de Saint-Etienne, 1986, 163p.

[LALLICH, 90] Lallich-Boidin G., Henneron G., Palermiti R. *.Analyse du français: Achèvement et implémentation de l'analyseur morpho-syntaxique*, Grenoble : Les cahiers du CRISS N°16, Université des Sciences Sociales de Grenoble, Nov. 1990, 123p.

[LAMROUS, 97] Sid Ahmed Lamrous, Philippe Trigano. Organisation des bases documentaires vers une exploitation optimale. *Document numérique.* 1997, Volume 1 – n°4, p.459-481.

[LANCASTER, 91] LANCASTER Frederic W. *.Indexing and Abstracting in Theory and Practice.* London : Library Association Publishing Ltd., 1991, 328 p.

[LANCASTER, 93] F.W. Lancaster, A. J. Warner. *Information Retrieval Today. chapter 5 : Subject Analysis and Representation.* Ed. Information Resources Press, 1993, p.79-88.

[LANDAUER, 91] Landauer T.K., Littman M. .A statistical Method for language-independent representation of the topical content of the text segments. *Textes des Communications IDT'91,* Avignon 1991, Paris : Ed. adbs & anrt, p.77-85.

[LARDY, 96] Jean-Pierre Lardy. *Recherche d'information dans Internet : Outils et méthodes*. ADBS Editions (2nd éditions), 1996, 100p.

[LAROUK, 93a] O. Larouk, R. Bouché. *Apports de la linguistique et des logiques dans la conception d'interface de bases de données textuelles.* Pluridisciplinarité dans les sciences cognitives : Textes réunis par O. Bousaïd *alii*, Paris : Editions Hermès, 1993, pp.142-160

[LAROUK, 93b] LAROUK Omar. *Extraction de connaissances à partir de documents textuels : traitement automatique de la coordination (connecteurs et ponctuation).* Thèse de doctorat: Université Claude Bernard Lyon 1, 1993. 290p.

[LAROUK, 93c] LAROUK Omar. Linguistico-statistical approach and logic applied in documentary system. *Proceedings of 8th SIGAPP Symposium on Applied Computing*, Indianapolis, 1993, p.737-744.

[LASSALLE, 93] LASSALLE E. .Telmi: a reusable information retrieval system and its applications. in *ASLIB Proceedings*, 1993, vol. 45, n°. 5, p.144-148.

[LAYAIDA, 96] Nabil Layaïda, loay sabry-Ismaïl. MADEUS : un modèle de documents multimédia structurés. *Techniques et science informatiques.* 1996, Vol.15, n°9, p.1227-1257.

[LE BORGNE, 96] Jean-Alain Le Borgne. *PatSco : extraction adaptative d'informations et structuration automatique de documents en texte libre (draft).* Rapport de Recherche, Département IA : Université Paris 8, 1996, 17p.
[On-line]- <URL : http://www.ai.univ-paris8.fr/~jalb/patsco/report1.html >

[LE CROSNIER, 97] Le Crosnier Hervé. Les bibliothèques numériques. *Communications Forum initiatives : inforoutes et technologies de l'information,* 26 octobre 1997 : Chine.
[On-line]- <URL : http://www.admiroutes.asso.fr/espace/acces/hanoi.html >

[LE GUERN, 82] LE GUERN Michel. Les descripteurs d'un système documentaire : essai de définition. In. : Bès, G.C., Fauchère, P.M., Lagueunière, F. *Actes du Colloque Traitement automatique des langues naturelles et systèmes documentaires :* Université Clermont Ferrand, 1982, p.163-173.

[LE GUERN, 89] Le Guern M. Sur les relations entre terminologie et lexique. *in actes du colloque: les terminologies spécialisés - Approches quantitatives et logico-sémantique*, et *Meta* Vol.34, No.3., sept. 89.

[LE GUERN, 91] LE GUERN Michel. Un analyseur morpho-syntaxique pour l'indexation automatique. *Le Français Moderne*. Juin, 1991, tome LIX, n°. 1, p.22-35.

[LE GUERN, 94a] LE GUERN Michel. *Traitement automatique et variation linguistique : la syntaxe des titres.* Opérateurs et constructions syntaxiques : Evolutions des marques et des distributions du XV-ème siècle, Paris : Presses de l'Ecole Normale Supérieure, 1994, p. 75-81.

[LE GUERN, 94b] LE GUERN Michel. *Parties du discours et catégories morphologiques en analyse*

automatique. Les Classes de Mots, Lyon : Presses Universitaires de Lyon, 1994, p. 207-215.

[LE LOARER, 89] Pierre Le Loarer, Bernard Normier. Nouveaux traitements de l'information pour l'entreprise. in *Textes des Communications IDT'89,* 1989, Paris : Ed. adbs & anrt, p.149-155.

[LEACH, 76] E. Leach. *Culture and Communication : The Logic by which Symbols are constructed*. UK : Ed. Cambridge, 1976.

[LEFEBVRE, 94] Alain Lefebvre. L'architecture Client-Serveur: aspects techniques, enjeux stratégiques. Paris : Armond Collin Editeur, 1994, 253p.

[LEFEBVRE, 97] Patrick Lefebvre, Eric Bomal, Michel De Heaulme. Une méthode et un algorithme pour compléter et abstraire une spécification orientée-objet. in *Actes du colloque langages et méthodes à objets (LMO'97)*, Roscoff France, 22-24 octobre 1997, ed. HERMÈS Paris, p.79-94.

[LELU, 97] A. Lelu, A-G. Tisseau-Pirot, A. Adnani. Cartographie de corpus textuels évolutifs : les options ergonomiques de NEURONAV+. *Hypertextes et hypermédias*, 1997, Vol.1,n°1, p.23-55.

[LESAGE, 96a] Olivier LESAGE, Julien Maltaverne. D'une planète à l'autre: Dites GEIDE ouvre-toi !. in *La tribune des industries de la langue et de l'information électronique*, Juillet 1996, N°.20-21-22, p.70-77.

[LESAGE, 96b] Olivier LESAGE, Julien Maltaverne. Le traitement avancé de l'information pour quoi faire ?. in *La tribune des industries de la langue et de l'information électronique*, Juillet 1996, N°.20-21-22, p.77-81

[LESPINASSE, 2001] Karine Lespinasse, Bruno Bachimont. Is Peritext a Key for Audiovisual Documents? The Use of Texts Describing Television Programs to Assist Indexing. in *CICLing 2001, Electronic Edition* (Springer LINK), 2001, p.505-506.
[On-line]- <URL : http://link.springer.de/link/service/series/0558/bibs/2004/20040505.htm >

[LEVINE, 96] Levine John R., Baroudi Carol. *Internet : les fondamentaux.* International THOMSON Publishing France : Paris, 1996, 976p.

[LEVY, 95] David M. Levy, Catherine C. Marshall. Going Digital: a look at assumptions underlying digital libraries. *Communications of the ACM,* April 1995, Vol.38, N°.4, p.77-85.

[LEVY-ABEGNOLI, 96] Thierry Levy-Abégnoli. . La recherche en texte intégral. in 01 Réseaux, N°. 30, pp.64-69, Septembre 1996.

[LEZAUM, 96] Jean-Louis Lezaum. Les enjeux d'internet dans l'entreprise. in *PC Expert*, octobre 1996, p.163-178.

[LIEBERMAN, 72] P. Lieberman et *alii*. .*The speech of Primates*. The Hague Ed., 1972.

[LIN, 95] D. Lin. Description of the PIE System as Used for MUC-6. In *Proceedings of the Sixth Conference on Message Understanding (MUC-6)*, Columbia, Maryland. 1995, University of Manitoba.
[On-line]- <URL : http://www.cs.umanitoba.ca/~lindek/publication.htm >

[LIPPMAN, 96] A. Lippman, R. Kermode. Media Banks : Entertainment and the internet. *IBM systems journal,* 1996,Vol.35, n°3-4, p.272-391.

[LIPPOLD, 97] B. Lippold, R. Kozlowska-Heuchin, T. Poibeau. NTK.FOCUS : Un outil d'aide à la rédaction des documents techniques. *Actes JST Francil 1997*, Editions AUPELF-UREF, (Avignon) France, p.369-373.

[LOINTIER, 96] Cécile Lointier. *Les bibliothèques électroniques.* Rapport de recherche bibliographique DESS Informatique Documentaire, ENNSIB de Lyon, 1996, 23p.

[LUSTIÈRE, 99] Colette Lustière. *Les différents types de notices dans BASINA: les étapes du traitement documentaire à la vidéothèque d'actualités INA*. Rapport technique interne INA: 20/10/1999, 13p.

[MAHMOUDI, 95] S.M. Mahmoudi. Traitement automatique des documents et le contenu du texte. *Informatique Documentaire* : Bulletin du centre de hautes études internationales d'informatique documentaire, N°4, 4eme trimestre 1995, Ed. CID Paris, p.53-65.

[MANIEZ, 77] Maniez J. *.Le rôle de la syntaxe dans les systèmes de recherche documentaire - Tome1: Aspects linguistiques, et - Tome 2: Etude critique de quelques SRD*. IUT de Dijon-Département carrières de l'information, 1976-77, 184p. et 182p.

[MANIEZ, 87] Maniez J. *.Les langages documentaires et classificatoires : conception, construction et utilisation dans les systèmes documentaires.* Paris : Editions d'Organisation. (Coll. Système d'information et de documentation), 1987.

[MANIEZ, 88] Maniez J. .Relationship in thesauri : some critical remarks. *International Classification,* 1988, Vol. 15, n° 3, p.133-138.

[MANIEZ, 91] Maniez J., de Grolier E. .A decade of research in classification. *International Classification,* 1991, Vol. 18, n° 2, p.73-77.

[MANIEZ, 93] Jacques Maniez. L'évolution des langages documentaires. *Documentaliste et Sciences de l'information,* 1993, vol.30, n°4-5, p.254-259.

[MARCHIONINI, 95a] Gary Marchionini. *Information seeking in electronic environments*. Cambridge University Press, 1995, 224p.

[MARCHIONNI, 95b] Gary Marchionni, Hermann Maurer. The roles of digital libraries in teaching and learning. *Communications of the ACM*, April 1995, Vol. 38, N°.4, p.67-75.

[MARET, 94] P. Maret, J-M. Pinon, D. Martin. Capitalisation of consultants' experience in document drafting. *Conference Proceedings RIAO 1994*, Printed by CID Paris France, p.113-118.

[METZGER, 85] J-P. Metzger. Bases de données textuelles et analyse morphosyntaxique. in *Textes des Communications IDT'85*, Versailles 12-14 Juin 1985, Ed. adbs & anrt Paris, p.33-38

[METZGER, 88] METZGER Jean-Paul. *Syntagmes Nominaux et Information Textuelle : reconnaissance automatique et représentation.* Thèse de Doctorat d'Etat en Sciences : Université Claude Bernard – Lyon 1, 1988, 324 p.

[MICH, 96] L. Mich. NL-OOPS : from natural language to object oriented requirements using the natural language processing system LOLITA. in *Natural Language Engineering*, 1996, vol.2/n°2, Cambrige University Press, p.161-187.

[MICHEL, 96] Jean-Luc Michel. Schématisation et cognition sur les réseaux hypertextuels. *in Revue de bibliologie- Schéma et schématisation*, N°. 44, pp. 28-33, 1996.

[MIGNOT, 94] Yvonne Mignot-Lefebvre. Vidéo et multimédia. *Actes du Séminaire : Ecrit Image Oral et Nouvelles technologies*, 1993-1994, Publications Université Paris 7, p.79-92

[MOHRI, 96] M. Mohri. On some applications of finite-state automata theoryto natural language processing. in *Natural Language Engineering*, 1996, vol.2, n°1, Cambrige university press, p.61-80.

[MORIN, 99] E. Morin, C. Jacquemin. Expansion automatique de thésaurus à partir de corpus. in *Actes Ingénierie des connaissances GRAC/AFIA (IC'99),* Palaiseau, p.97-106

[MOULIN, 91] B. Moulin, D. Rousseau. Structuration d'une base de connaissances à partir des informations extraites de textes prescriptifs. in *Actes du colloque ILN'91*, Janvier 1991, Nantes-France, Ed. Université de Nantes, p.1-18.

[MULDER, 72] J.W.F. Mulder, S.G.J. Harvey. *Theory of the Linguistic Sign*. Janua Linguarum, Series Minor : 136, La Haye, 1972.

[MUSGRAVE, 96] J.F. Musgrave, M.R. Cooper. Experiments in digital graphic design. *IBM systems journal,* Vol.35, n°3-4, 1996, p.499-513.

[MUSTAFA, 89] Mustafa Elhadi W. *.La terminilogie arabe des télécommunications : Faits de variations*. Thèse de Doctorat en Science du langage : Université Lyon2, 1989.

[MUSTAPHA, 92] Mustafa Elhadi W. .La contribution de la terminologie à la conception théorique des langages documentaires et à l'indexation de documents. *Meta,* 1992,vol. XXXVII, n°3,

p.465-473.

[NAPLI, 2000] A. Napoli, J. Euzenat, R. Ducournau. La représentation des connaissances par objets. *Technique et Science Informatiques*, 2000, Vol.19, n°1-2-3, p.387-394.

[NAPOLI, 92] A. Napoli. *Représentation à objets et raisonnement par classification en intelligence artificielle.* Thèse d'Etat : Université de Nancy 1, France, 1992.

[NAPOLI, 94] A. Napoli. Catégorisation raisonnement par classification et raisonnement à partir de cas. in *Actes des 9ème Journées Française de l'Apprentissage*, Strasbourg : France, p.E1-14.

[NASTAR, 97] Chahab Nastar. Indexation d'images par le contenu : un Etat de l'Art. in *Actes Colloque CORESA*, *Issy-les-Moulineaux*, 26-27 mars 1997, p.8

[On-line]- <URL : http://www-rocq.inria.fr/~nastar/publications.html >

[NIBLET, 88] T. Niblet. Astudy of generalization in logic programs. *in EWSL'88, Galsgow England*, p.131-146.

[NIE, 97] J-Y. Nie, J-P. Chevallet, Y. Chiaramella. Vers la recherche d'informations à base de termes. *Actes JST Francil 1997*, Editions AUPELF-UREF, (Avignon) France, 1997, p.119-125.

[OTMAN, 92] OTMAN Gabriel. .Des ambitions et des performances d'un système de dépouillement terminologique assisté par ordinateur. *La Banque des Mots, Numéro spécial CTN 1991*, CILF, Paris, 1992.

[OZKARAHAN, 95] Esen Ozkarahan. Multimedia Document Retrieval. in *Information Processing & Management,* vol.31, n°1, 1995, Elsevier Science Ltd Printed G.B., p.113-131.

[PACK, 95] Thomas Pack. Digital Circles of Knowledge. in *DATABASE,* december 1995, p.14-23.

[PÊCHEUX, 69] Pêcheux M. .Analyse automatique du discours. Paris : DUNOD – Science du Comportement, Collection dirigée par F. Bresson et M. de Montmollin, 1969, 140p.

[PELEATON, 2000] R.A. Peleaton, J-C. Chappelier, M. Rajman. Automated Information Extraction out of Classified Advertisements, in *5th International Conference on Applications of Natural Language to Information Systems (NLDB'2000),* Versailles France, june 2000

[On-line]- <URL : http://liawww.epfl.ch/~chaps/index.html >

[PELOU, 85] Pierre Pelou, Alain Vuillemin. *Les nouvelles technologies de la documentation et de l'information*, Ed. la documentation française – Paris, 1985.

[PENTLAND, 94] A. Pentland, R. Picard, G. Davenport, K. Haase. Video and image semantics : advanced tools for telecommunications. *Conference Proceedings RIAO 1994*, Printed by CID Paris France, p.3-12.

[PEREIRA, 80] F.C.N. Pereira, D.H.D. Warren. Definite Clause Grammars for Language Analysis – a survey of the formalism and a comparison with Augmented Transition Networks. *Artificial Intelligence Journal.*, By North-Holland Publishing Company, 1980, vol.13, p. 231-278.

[PERRIN, 95] Dominique Perrin. Les débuts de la théorie des automates. *Technique et Science Informatiques*, 1995,Vol.14, n°4, p.409-433.

[PERRON, 89a] PERRON Jean. "Termino : un système de dépouillement terminologique", Terminogramme Ndeg.54, 1989.

[PERRON, 89b] PERRON Jean. ."Un système de dépouillement terminologique". *Revue ICO,* Vol. 2, Ndeg.5, 1989.

[PICARD, 96] R.W. Picard. A society of models for video and image libraries. *IBM systems journal*, 1996, Vol.35, n°3-4, p.292-312.

[PIGOT, 96a] Thierry PIGOT. INTRANET: une déferlante annoncée. in *Réseaux & Télécoms*, Octobre 1996, p. 6-9.

[PIGOT, 96b] Thierry PIGOT. Trois dimensions-trois conception. in *Réseaux & Télécoms*, Octobre

1996, p. 22-25.

[PINON, 96] Jean-Marie Pinon. *Projet SEMUSDI : Serveur de documents Multimédia en Sciences de l'Ingénieur.* Rapport de Présentation Technique : insa de Lyon, Juillet 1996, 15*p.*

[PLANTE, 88] PLANTE Pierre. .*Le dépistage automatique des termes.* Terminogramme Ndeg.46, 1988.

[On-line]- <URL : http://www.uhb.fr/langues/balneo/html/inv9459.htm#Heading573>

[PLOTKIN, 70] G. Plotkin. A note on inductive generalization. *in B. Meltzer and D. Michie, editors, Machine Intelligence,* New York : Elsevier, p.101-121.

[POEHL, 92] Karsten Poehl. Le syntagme nominal comme outil de l'indexation automatique : le cas de la langue allemande. *Les cahiers des sciences de l'information et de la communication,* n°2, Février 1992, Publications Universités : Grenoble 2 et 3 et Lyon 2 et 3, p. 3-21.

[PONCHET, 97] Marc Ponchet. Le projet de Centre d'Information Scientifique Multimédia (CISM) de CEA. *Document numérique,*1997, Volume. 1, n°3, p.351-361.

[PORTE, 96] Olivier Porte. Les techniques de spécifications formelles. *Le Micro Bulletin*, Novembre/Décembre 1996, n°6, p.56-93.

[POULAIN, 96] Gérard Poulain. *Métaphore et Multimédia : concepts et applications*, Paris : La documentation Française, 1996.

[PRIÉ, 2000] Yannick Prié. Sur la piste de l'indexation conceptuelle de documents : une approche par l'annotation. *Document Numérique*, 2000, Vol.4, n° 1-2, p.11-35.

[PRINCE, 97] V. Prince, S. Ferrari. Création et extension automatiques de dictionnaires terminologiques multi-lingues spécialisés. *Actes JST Francil 1997,* Editions AUPELF-UREF, (Avignon) France, p. 555-560.

[PUGEAULT, 96] F. Pugeault, M-G. Monteil. *Une étude pour l'extraction d'index structurés à la DER.* Rapport de Recherche : IRIT & EDF, Université P. Sabatier (Toulouse) France, 1996, p.165-175.

[QUELLEC, 97] B. Le Quellec, J-L. Vuldy. Génération automatique d'hypermédias à caractère technique. *Hypertextes et hypermédias,* 1997, Vol.1, n°1, p.103-113.

[RADA, 95] Roy Rada. *Interactive Media*. Ed. Springer-Verlag – New York, 1995.

[RAJMAN, 97] M. Rajman, R . Besançon. Text Mining : Natural Language Techniques and Text Mining Applications. in *IFIP 1997*, Published by Chapman & Hall.

[On-line]- <URL : http://liawww.epfl.ch/ >

[RAJMAN, 98a] M. Rajman, L. Lebart. Similarités pour données textuelles. in Proceedings 4th International Conference on Statistical Analyse on Textual Data (JADT'98) , feb. 1998, Nice France.

[RAJMAN, 98b] M. Rajman, R. Besançon. Text Mining – Knowledge extraction from unstructured textual data. in *6th Conference on International Federation of Classification Societies*, July. 1998, Roma-Italy, p.473-480.

[RAMBURRUN, 85] Ramburrun Mooneswar. Implémentation en Prolog-Foll d'un analyseur morpho-syntaxique du francçais écrit. Thèse de Doctorat en Science du langage : Université Lyon2, 1985. 166p.

[RANJARD, 95] Sophie Ranjard. Indexer et résumer, pourquoi et comment ? . *Archimag,* décembre-janvier 1995, n°80, p.41-43.

[RESCHE, 89] Danielle Resche. *L'indexation de l'image fixe.* Note de synthèse DESS Informatique Documentaire : Université Lyon 1, 1989, 18p.

[RHISSASSI, 97] Habib Rhissassi, Alain Lelu. Vers un environnement complet d'indexation automatique. in *Hypertextes et Hypermédias.* 1997, Vol.1, n°2-3-4, p.225-236.

[On-line]- <URL : http://hypermedia.univ-paris8.fr/paragraphe/articles/c1rhiss.htm >

[RICHARD, 96] Jean-Philippe Richard. La dynamique du WEB. in *Réseaux & Télécoms*, Octobre 1996, p. 26-29.

[RICHAUDEAU, 99] François Richaudeau. Les complexités de la communication par l'écrit. *Revue de Bibliologie : Schéma et Schématisation,* 1999, n°45, p.27-34.

[RISLER, 99] Robert RISLER. La visualisation des données par la chromique. *Revue de Bibliologie : Schéma et Schématisation*, 1999, n°45, p.76-81.

[RIVIER, 90] Alexis Rivier. Construction des langages d'indexation : Aspects théoriques. *Documentaliste*, novembre-décembre 1990, vol. 27, n°6, p.263-274.

[ROBINSON, 95] Robinson Andrew. *The Story of Writing: Alphabets, Hieroglyphs and Pictograms*. New York: Thames and Hudson Inc., 1995.

[ROMER, 95] D.M. Romer, E.Kodak. Image and Multimedia Retrieval. 1995.

[On-line]- <URL : http://www.ahip.getty.edu/agenda/image.html>

[ROUAULT, 88] Rouault J. .Apport Contrainte et limites du langage dans le traitement automatique des langues. in *Colloque Fribourg, Suisse,* Mars 1988, 23p.

[ROUSSEAU, 94] Bertrand Rousseau, Mario Ruggier. Writing documents for paper and WWW: a strategy based on FrameMaker and WebMaker. in *Computer Networks and ISDN Systems,* , 1994, N° 27, p.205-214.

[ROUSSEAU-HANS, 98] F. Rousseau-Hans. L'analyse de corpus d'information comme support de la veille stratégique. in *Document Numérique,* 1998, Vol.2, n°2, p.177-202.

[SABAH, 90] Gérard Sabah. *L'intelligence artificielle et la langage – Représentation des Connaissances.* Editions Hermès (2nd édition), Paris, 1990.

[SAINT-DIZIER, 85] Patrick Saint-Dizier. *An approach to natural language semantics in logic programming*. Publication Interne N° 251-IRISA, Rennes, Mars 1985 , 34p.

[SALTON, 73] Salton G., Wong A., Yu C.T. .Automatic indexing using term precision measurements. *Information Processing and Management*, 1973, vol.12, p.43-51.

[SALTON, 83] SALTON Gerard, MCGILL Michael J., *Introduction to modern information retrieval*. New York : Mcgraw-Hill Book Company (Computer Science), 1983, 448p.

[SALTON, 86a] Salton G. .On the use of term associations. *in Automatic information retrieval*, *COLLING'88 12th International Conference on Computationnal Linguistics*, Bonn 1986, p.380-386.

[SALTON, 86b] Salton G. .Another look at automatic text-retrieval systems. *Communications of the ACM*, 1986, vol. 29 (7), p.648-656.

[SALTON, 89] SALTON Gerard, *Automatic Text Processing : The Transformation, Analysis and Retrieval of Information by Computer*. Massachusetts : Addison-Wesley Publishing Co. (Computer Science), 1989. 530 p.

[SANDFORD, 97] E. Sandford, J. Chauché. Une désambiguïsation sémantique pour affiner les résultats d'une interrogation d'une base de données textuelles. *Actes JST Francil 1997*, Editions AUPELF-UREF, (Avignon) France, 1997, p.89-93.

[SCHILLER, 96] Anne Schiller. Multilingual Finite-State Noun Phrase Extraction. in *ECAI 96. Workshop on Extended Finite-State Models of Language*, Edited by A. Kornai, Published by John Wiley & Sons Ltd., 1996

[SCHMUCK, 95] Claudine Schmuck. *Introduction au multimédia: technologies et marchés*. Publication AFNOR, Paris, 1995, 233p.

[SCHÜTZENBERGER , 77] M. P. Schützenberger. Sur une variante des fonctions séquentielles. *Theoretical Computer Science*, 4(1):47-57, February 1977.

[SEVIGNY, 96] Martin Sévigny, Yves Marcoux. Construction et évaluation d'un prototype d'interface-utilisateurs pour l'interrogation de bases de documents structurées. *Canadian Journal of Information and Library Science*, Vol.21, No.3-4, September-December 1996, p.59-77.

[SIDHOM, 2000] Sidhom Sahbi .The audiovisual : content analysis and automatic filtering of textual information. *Proceedings of World Multiconference on Systemics, Cybernetics and Informatics,* July 23-26-2000. Orlando-Florida USA, Volume X, p.520-525.

[SIDHOM, 98] Sidhom Sahbi. Automatic indexing of multimedia documents based on the extraction of nominal phrases. *Proceedings of 5th ISKO Conference,* 25-29 august 1998 Lille, France, Ed. Ergon Verlag.

[SIDHOM, 99a] Sahbi Sidhom, Mohamed Hassoun, Richard Bouché. Cognitive grammar for indexing and writing. *ISKO-España Conference Proceedings*, 22-24 april 1999 Granada, p.11-16.

[SIDHOM, 99b] Sahbi Sidhom, Mohamed Hassoun, Richard Bouché, Colette Lustière, Daniel Gegez. Multimédia et exploitation textuelle pour un modèle d'indexation automatique. *ISKO'99 Proceedings,* Lyon France, 21-22 Octobre 1999.

[SILBERZTEIN, 2001] Max Silberztein. [On-line] INTEX System – Overview. 1998-2001,

[On-line]- <URL : http://ladl.univ-mlv.fr/INTEX/index.html >

[SMEATON, 89] SMEATON Alan F. Information retrieval and natural language processing. in.: *Informatics 10: prospects for intelligent retrieval: proceedings of a conference jointly sponsored by ASLIB.* Cambridge : University of York, 1989, 21-23 Mars, p.1-14.

[SMEATON, 91a] SMEATON Alan F., SHERIDAN Paraic. Using Morpho-Syntaxique Language Analysis in Phrase Matching. *RIAO 9 Proceedings : Recherche d'Information Assistée par Ordinateur.* Barcelona, 1991, vol. 1, p.414- 430.

[SMEATON, 91b] SMEATON Alan F. .Prospects for intelligent, languaged-based information retrieval. *Online Review*, 1991, vol. 15, n°.6, p.373-382.

[SRPOVA, 95] Milena Srpova. La traduction, confrontation de deux expériences cognitives. in *Intellectica*, 1995, vol.1, n°.20, p.157-170.

[STILES, 61] Stiles H.F. .The association factor in information retrieval. *Journal of the ACM*, 1961, vol. 8, p.271-279.

[SUGIYAMA, 96] Kenji Sugiyama. Un logiciel pour la constitution automatique de bases de connaissances. *La lettre d'information d'ELRA*, Décembre 1996, pp.16-17.

[SYSTEX, 92] Articles Pressbook-SYSTEX. *Indexation automatique et interrogation de bases de données textuelles en langage naturel.* in Tribune des industries de la langue, *Novembre 1992, n° 10, 4p.*

[TCHOBANOV, 95] Tchobanov Atanas. *Le modèle syntaxe sémantique : analyse de la phrase simple du français pour le TAL.* Mémoire de Maîtrise en Science du Langage : Université Paris X-Nanterre, 1995, p.120.

[THIL, 96] Jérôme Thil. Outils intelligents de recherche d'informations : mythe ou réalité?. in *Technologies Internationales*, Juillet-Aout 1996, N°. 26, p.7-10.

[THIVOLLE, 98] Laurence Thivolle. Apport de la méthode d'analyse iconographique d'Erwin Panofsky pour l'analyse des images. in *Canadian Journal of Information and library Science,* Avril-Juillet 1998, Vol.23, n°1-2, p.31-49.

[TOPIC, 94] TOPIC-VERITY. *TOPIC: la recherche documentaire par concept*. Rapport technique copyright VERITY France, Fév. 1994, 17p.

[TOUSSAINT, 97] Yannick Toussaint. L'analyse de l'information par la construction à partir de textes d'une base de connaissance partielle : le projet ILIAD. *Actes JST Francil 1997,* Editions AUPELF-UREF, (Avignon) France, 1997, p. 517-521.

[TREMBLAY, 86] Diane Tremblay. *Représentation sémantique et syntaxique de termes dans les dictionnaires électroniques*. Thèse de Doctorat : Université Paris 8, 1986, 331p.

[TURNER, 96] James M. Turner, Catherine Lavallée, Abdeljalil Zyati. Le développement d'une méthodologie pour la construction de systèmes d'information multimédias. *Documentation et bibliothèques,* Juillet-Septembre 1996, p.119-125.

[TURTLE, 91] Howard R. Turtle, W. Bruce Croft. Evaluation of an Inference Network-Based

Retrieval Model. *TOIS 9(3): 187-222 (1991)*

[VAN HOUCKE, 94] Van Houcke Christian. *Le multimédia en entreprise.* HERMÈS Editions,1994, 188p.

[VAN SLYPE, 87] VAN SLYPE George. *Les langages d'indexation : conception, construction et utilisation dans les systèmes documentaires.* Paris : Les éditions d'Organisation, 1987. 277 p.

[VAPILLON, 97] Jérôme Vapillon. Un atelier de Génie Linguistique pour l'analyse syntaxique fondée sur le formalisme LFG. *Actes JST Francil 1997,* Editions AUPELF-UREF, (Avignon) France, 1997, p. 375-381.

[VARELA, 98] Francisco Varela. Le cerveau n'est pas un ordinateur. in *La recherche,* Avril 1998, n°308, p.109-113.

[VEENEMA, 96] Fred Veenema. L'indexation automatique – oui ou non ?. *Canadian Journal of Information and Library Science*, juillet 1996, Vol.21, n°2, p.1-22.

[VETTRAINO, 99] Marie-Claude VETTRAINO-SOULARD. Vingt ans de recherches iconologiques. *Revue de Bibliologie : Schéma et Schématisation*, 1999, n°45, pp.69-75.

[VIEIRA, 96] Lise Vieira. Les fonctions métaphoriques du Multimédia. *Communication et langages* 1996, n°109 – 3eme trimestre, Editions Retz, Paris, p.98-111.

[VINESSE, 97] Jérôme Vinesse. Acquisition automatique de données lexicales à partir de corpus textuelles. *Ecole d'été CNET (5, Trégastel 1995) – Traitement des langues naturelles*, Editions CNET, pp. V.1-13.

[WARREN, 82] David Scott Warren, Joyce Friedman. Using Semantics in Non-Context-Free Parsing of Montague Grammar. *American Journal of Computational Linguistics 8*(3-4): 123-138 (1982)

[WEHRLI, 91] Eric Wehrli. *Stratégie d'analyse et structures de données.* Notes techniques 1991/n°7, Laboratoire d'analyse et de technologie du langage – université de Genève, 1991, 17p.

[WERMTER, 92] Stefan Wermter. Learning Natural Language Filtering Under Noisy Conditions. *Proceedings of the Tenth IEEE Conference on Artificial Intelligence for Applications*. San Antonio USA., 1994, p.215-221

[On-line]- <URL : http://osiris.sunderland.ac.uk/~cs0stw/ >,
<URL : http://www.informatik.uni-hamburg.de/Arbeitsbereiche/NATS/staff/ >

[WERMTER, 95] Weber V., Wermter S. .Artificial Neural Networks for Repairing Language. in *Proceedings of the 8th International Conference on Neural Networks and their Applications.* Marseilles, FRANCE. 1995.

[On-line]- <URL : http://www.informatik.uni-hamburg.de/Arbeitsbereiche/NATS/staff/ >

[WIEGANDT, 94] Caroline Wiegandt, Eva Dauphin. La caractérisation lexicale de corpus dans le cadre d'applications documentaires. in *Textes des Communications IDT'94*, 1994, Paris, , Ed. adbs & anrt Paris, p.94-96.

[WINOGRAD, 83] Terry Winograd. *Language as a cognitive process – Syntax.* Editions Addition Weslery Publish company, Stanford University usa., 1983.

[WOODS, 70] William A. Woods. Transition Network Grammars for Natural Language Analysis. *in Communication ACM.*, October 1970, Vol.13, n°10, p.591-606.

[WOODS, 73] William A. Woods. *An Experimental Parsing System for Transition Network Grammars.* in Natural Language Processing, Randall Rustin, ed. New York : Algorithmics Press, 1973.

[WOODS, 80] William A. Woods. Cascaded ATN Grammars. in *American Journal of Computational Linguistics*, January-March 1980, vol.6, n°1.

[WOODS, 86] William A. Woods. Transition Network Grammars for Natural Language Analysis. in : Natural Languge Processing, San Mateo : Morgan Kaufmann, 1986.

[WOODS, 97] William A. Woods. *Conceptual Indexing : a better way to organize knowledge.*

Technical Report SMLI TR-97-61 : SUN Micosystems, Lab. Mountain View Canada, April 1997

[On-line]- <URL : http://sun.com/research/techrep/1997/ >

[WOODS, 98a] William A. Woods, J. Ambroziak. Natural language technology in precision content retrieval. in *Proceedings NLP+IA'98,* August 1998, Moncton - New Brunswick CANADA

[WOODS, 98b] William A. Woods. Knowledge Management needs effective search technology. in *SUN Journal*, March 1998

[On-line]- <URL : http://www.sun.com/sun-journal/ >

[WYLLYS, 62] Wyllys R.E., *Automatic analysis of contents of documents- Part 2: document searches and condensed representation*, Rapport interne FN-6170, System Development Corporation, 1962.

[YU, 77] Yu C.T., Salton G. .Effective information retrieval using term accuracy. *Communications of the ACM*, 1977, vol.20., p.135-142.

[ZAHARIN, 88] Yosoff Zaharin. Towards an analyser (parser) in a machine translation system based on ideas from experts systems. in *Comput. Intell.,* 1988, N°4, p.180-191.

[ZIGHED, 2000] Zighed D.A., Rakotomalala R., *Graphes d'induction : Apprentissage et Data Mining.* Hermès Science Publications : Paris, 2000, 475p.

[ZINGLÉ, 97] H. Zinglé. Développement et mise en oeuvre de ressources linguistiques avec la Zstation. *Actes JST Francil 1997,* Editions AUPELF-UREF, (Avignon) France, 1997, p. 321-325.

[ŽINKIN, 62] ŽINKIN N.I., Four Communicative Systems and Four Languages. Word n°18 vol.1-2, 1962

[ZWEIGENBAUM, 94] ZWEIGENBAUM P. .MENELAS: an access system for medical records using natural language. *Computer Methods and Programs in Biomedicine*. 1994, vol. 45, n°. 1-2. p. 117-120.

Résumé

La contribution de ce travail de thèse s'inscrit au sein d'un domaine multidisciplinaire regroupant le traitement automatique du langage naturel, l'indexation dans un système d'information documentaire et l'organisation des connaissances autour de l'information écrite. Sa particularité consiste en la mise à disposition d'outils pour le traitement automatique de l'information.
L'objectif est de construire *une Plate-forme d'analyse morpho-syntaxique* pour l'indexation automatique et la recherche d'information. Elle est composée d'un noyau d'indexation automatique (*processus d'indexation*) qui utilise le modèle des syntagmes nominaux comme descripteurs de l'information textuelle. Ces syntagmes sont organisés selon une approche Logique Intensionnelle/Extensionnelle (*processus de classification des connaissances*) qui permet d'ordonner les objets d'une classe et de distinguer les classes de connaissances. A la base de cette dernière propriété, nous construisons notre approche pour la recherche d'information (*processus de recherche d'information*).
Cette Plate-forme d'analyse dans sa logique de fonctionnement sera un outil d'investigation orienté vers l'organisation et la gestion des connaissances écrites.
Dans notre recherche, cet aspect sur l'organisation des connaissances a été conduit dans le but de faire émerger les propriétés linguistiques et le traitement du langage dans une pratique expérimentale sur l'indexation automatique documentaire. Nous avons montré la nécessité de coordonner d'autres sources et stratégies dans l'exploration de ces propriétés. Il s'agit du mode de raisonnement et de la technique d'exploitation des objets du discours spécifiques à la gestion des connaissances (comme étape préalable à la recherche d'information).
Ces deux derniers aspects (mode et technique) intégrés dans le processus de la présentation et de l'organisation du syntagme nominal offrent des scénarii pertinents pour la recherche d'informations.

Mots-clés

Plate-forme d'analyse morpho-syntaxique, organisation des connaissances, recherche d'information, indexation automatique, syntagme nominal, traitement automatique du langage naturel, programmation par objets, formalisme ATN.

Abstract [**Title=** Morpho-syntactic Analysis Platform for Automatic indexing and Information retrieval : from text-writing to knowledge management]

The contribution of this thesis work is registered within a multidisciplinary field gathering the natural language processing, the documentary indexing in information system and the knowledge organization about written information. Its characteristic consists of the availability of automatic tools for information processing.
The objective is to build a morpho-syntactic analysis Platform for the automatic indexing and the information retrieval. It is made up of an automatic indexing core (*indexing process*) which uses the noun phrases (or nominal syntagms) model like descriptors of textual information. These syntagms were organized according to Intensional/Extensional Logic approach (*knowledges classification process*). On the basis of the last property, we incorporate our approach for the information retrieval (*information retrieval process*).
This analysis Platform in its logic operating will be a tool for investigation directed towards the organization and the knowledge management of written information.
In our research, this aspect about the knowledge organization was led with an aim to make emerge the linguistic properties and the language processing in an experimental practice on the automatic documentary indexing. We showed the need to coordinate other sources and strategies in the browsing of these properties. It is the mode of reasoning and the operating technique of the speech objects specifically to the knowledge management (as a preliminary in the information retrieval).
These two last aspects (mode and technique) integrated in the process of the presentation and the organization of the noun phrase offer relevant scenarii for the information retrieval.

Keywords

Morpho-syntactic analysis platform, knowledge organization, information seeking, automatic indexing, noun phrase, natural language processing, object-oriented programming, ATN formalism.

Résumé étendu

La contribution de ce travail de thèse s'inscrit au sein d'un domaine multidisciplinaire regroupant le traitement automatique du langage naturel, l'indexation dans un système d'information documentaire et l'organisation des connaissances autour de l'information écrite. Sa particularité consiste en la mise à disposition d'outils pour le traitement automatique de l'information.

L'objectif est de construire *une Plate-forme d'analyse morpho-syntaxique* pour l'indexation automatique et la recherche d'information. Elle est composée d'un noyau d'indexation automatique (*processus d'indexation*) qui utilise le modèle des syntagmes nominaux comme descripteurs de l'information textuelle. Ces syntagmes sont organisés selon une approche Logique Intensionnelle/Extensionnelle (*processus de classification des connaissances*) qui permet d'ordonner les objets d'une classe et de distinguer les classes de connaissances. A la base de cette dernière propriété, nous construisons notre approche pour la recherche d'information (*processus de recherche d'information*).

Cette Plate-forme d'analyse dans sa logique de fonctionnement sera un outil d'investigation orienté vers l'organisation et la gestion des connaissances écrites.

Dans notre recherche, cet aspect sur l'organisation des connaissances a été conduit dans le but de faire émerger les propriétés linguistiques et le traitement du langage dans une pratique expérimentale sur l'indexation automatique documentaire. Nous avons montré la nécessité de coordonner d'autres sources et stratégies dans l'exploration de ces propriétés. Il s'agit du mode de raisonnement et de la technique d'exploitation des objets du discours spécifiques à la gestion des connaissances (comme étape préalable à la recherche d'information).
Ces deux derniers aspects (mode et technique) intégrés dans le processus de la présentation et de l'organisation du syntagme nominal offrent des scénarii pertinents pour la recherche d'informations.

Nous débutons ce mémoire, chapitre I, par une étude sur la connaissance et la transmission des savoirs.
Le chapitre II, sur l'état de l'art dans le domaine de l'indexation automatique, recouvre les activités documentaires en matière d'analyse de contenu et sa dynamique évolutive.
Le modèle linguistique du Groupe de recherche SYDO, objet du chapitre III, est le thème central dans la 're-formulation' de la problématique classique du descripteur en indexation documentaire.
Le chapitre IV exposera l'Extraction des Connaissances dans les processus d'analyse et d'indexation du contenu (audiovisuel) basée sur corpus.
Le chapitre V décrira une méthode de conception du noyau d'indexation automatique. L'architecture de l'application est basée sur les automates à transitions augmentées en cascade (C'ATN) de W. Woods (1970-80).
Au chapitre VI, nous naviguerons autour de l'objet 'descripteur' construit comme un syntagme nominal et relevant d'un principe d'organisation du discours. Cette dernière étape permet l'exploration des sources de connaissances selon les constructions logico-sémantiques des syntagmes nominaux.
Dans la partie annexe, nous complétons ce travail concernant la *plate-forme* par les interfaces Homme-machines et leurs modes d'exploitations. Le processus concerne la présentation de l'organisation des connaissances autour du syntagme nominal et offre des scénarii différents de recherche d'informations.