Université Lumière – Lyon 2

THÈSE présentée pour obtenir le titre de **DOCTEUR**
spécialité : Sciences de l'information et de la Communication
par

Hélio KURAMOTO
boursier du CNPq - Brasilia - Brésil

Proposition d'un Système de Recherche d'Information Assistée par Ordinateur

sous la direction de M. Michel LE GUERN
Février/1999

Table des matières

À ma mère Aylda, à mes filles Liliana et Luisa, et à ma femme Cristiane

Remerciements

- Je remercie vivement Monsieur Michel LE GUERN, professeur à l'Université Lumière, par ses enseignements, pour sa sagesse, pour sa patience, pour m'avoir confié ce travail et pour me faire partager son expérience dans la direction de cette thèse.
- Je remercie M. Richard BOUCHÉ, directeur de l'axe 1 du Centre d'Études et de Recherche en Sciences de l'Information (CERSI), pour l'accueil à l'ENSSIB et aussi au seins du CERSI.
- Je remercie Monsieur Omar LAROUK pour sa disponibilité, intérêt, ses aides et orientations qui m'ont permis de mener à bien ce travail.
- Je remercie Monsieur Emilio GIUSTI, professeur à l'Université Lumière, par sa disponibilité, ses enseignements et orientations sur la langue portugaise.
- Je remercie toute ma famille, mes amis et mes amies pour l'encouragement et l'appui qui m'ont donné pendant la durée de ces études.
- Je remercie CNPq pour m'accorder la bourse qui m'a permis de faire ces études.
- Je remercie Monsieur José RINCON FERREIRA, en tant que directeur de l'IBICT et aussi comme représentant de l'IBICT, de m'avoir accorder cette opportunité de perfectionnement de mon profil intellectuel et technique.
- Je remercie tous les professionnels, bibliothécaires, techniciens et administratifs, de l'ENSSIB pour leur accueil, leur aide et leur amitié.
- Je remercie infiniment mes amis Josette, Robert et Virginia pour l'encouragement, l'aide et l'appui qui m'ont donné pendant la durée de ces études.

Résumé

Nous proposons un modèle d'un système d'indexation et de recherche d'information afin de faire face aux difficultés rencontrées par les usagers lors de l'utilisation de tels systèmes. Nous en distinguons deux types de problèmes : la faible précision des résultats d'une recherche d'information et le manque de convivialité des interfaces de recherche d'information. Nous limitons notre champ d'étude aux systèmes de recherche d'information (SRI) qui portent sur des bases de données textes pleins (*full text*).

Suite à l'étude de ces difficultés nous sommes parvenus à une conclusion identique à celle proposé par le groupe SYDO : l'utilisation des syntagmes nominaux (SN) comme descripteur, en opposition à l'utilisation des mots couramment adoptés par les SRI classiques.

Il s'agissait alors d'examiner la faisabilité de cette proposition. Nous avons donc développé une maquette d'un SRI ainsi qu'une base de données construite à partir d'un corpus d'articles scientifiques en langue portugaise. Ensuite, nous avons exploité cette maquette à l'aide d'un thesaurus, cela nous a permis de mieux connaître le comportement des SN à l'intérieur d'une structure arborescente, ainsi que de l'interface de recherche d'information.

Pour conclure, nous avons établi un modèle de reconnaissance et d'extraction des SN en textes en langue portugaise.

Plus que simplement arriver à la conclusion pour la faisabilité de notre proposition, la démarche adoptée nous a montré que les connaissances obtenues dans la pratique d'extraction et d'indexation des SN, ainsi que dans le développement de la maquette ont été importants pour l'établissement du modèle de reconnaissance et d'extraction des SN.

Mots clés

Système de Recherche d'Information ; interface de recherche d'information ; indexation automatique ; syntagmes nominaux ; reconnaissance de syntagmes nominaux ; extraction de syntagmes nominaux ; traitement automatique du langage naturel.

Avant-propos

Le but initial de mes études en France était de proposer de construire un Système de Recherche d'Information (SRI) guidée par langage naturel. Ce but a été établi comme conséquence des expériences vécues dans mon métier. Je travaille dans l'IBICT (Institut Brésilien d'Information en Science et Technologie), depuis 1983. Cet institut est l'organisation brésilienne responsable pour la dissémination de l'information dans la communauté scientifique et technologique. Une des grandes difficultés trouvée par les chercheurs et par les techniciens en information était comment trouver l'information. En fait, les systèmes de recherche d'information n'offraient pas une interaction conviviale. Le manque de convivialité des interfaces de recherche d'information éloignait les usagers de l'information. Cela arrivait parce que d'une part il y avait une variété de langage de recherche (langage à commandes), chaque système avait un langage particulier. D'autre part, presque tous les systèmes utilisaient la logique booléenne comme moyen d'exprimer le besoin d'information de l'utilisateur. Il fallait donc bien connaître les langages de recherche d'information ainsi que la logique booléenne.

En 1987, nous avons coordonné un projet de construction d'un réseau d'information basé sur un réseau d'ordinateur en utilisant un langage commun de recherche d'information, laquelle toutes les institutions devraient utiliser. Ce langage a été basé sur une norme ISO appelé *Common Command Language (CCL)*. C'était une initiative dans le sens de faciliter l'accès à l'information aux usagers. Ce travail nous a motivé à faire des études ayant comme but la construction d'un système de recherche d'information avec l'interaction en langue naturelle. Un tel système serait beaucoup plus convivial que les systèmes de recherche d'information traditionnels. Pourtant ces études et l'analyse de la littérature spécialisée, concernant ce sujet, nous ont amenés à un autre problème, le manque de précision dans les résultats d'une recherche d'information dans les systèmes traditionnels. Il ne fallait pas donc seulement construire un système plus convivial, mais il fallait construire un nouveau système capable de donner aux usagers, de manière conviviale, les informations qu'ils attendaient.

Ainsi, pour la proposition de ce nouveau Système de Recherche d'Information, nous avons adopté les syntagmes nominaux comme moyen d'accès à l'information comme modèle d'indexation automatique. Mais, au lieu de construire d'abord une grammaire pour l'identification et pour l'extraction des syntagmes nominaux, nous avons d'abord vérifié la faisabilité de construire un SRI en utilisant les syntagmes nominaux comme descripteurs. Pour cela, nous avons tout d'abord constitué un corpus d'articles en langue portugaise et puis nous avons extrait les syntagmes nominaux de manière non automatique. C'est-à-dire manuellement, en utilisant une approche logico-sémantique. Par la suite, nous avons construit une maquette d'interface de recherche d'information et organisé, dans une structure arborescente, les syntagmes nominaux extraits. Nous avons donc développé une base de données d'articles en langue portugaise et aussi une interface de recherche d'information guidée par menu.

L'ensemble de syntagmes nominaux extraits du corpus d'articles a servi de base pour l'analyse et construction des règles de formation des syntagmes nominaux. Je ne suis pas un linguiste mais plutôt un professionnel de l'information et de l'informatique. Cette démarche a été très important car cette tâche initiale nous a permis de mieux connaître le comportement des syntagmes nominaux dans les textes en langue portugaise. Grâce à cette démarche, nous avons pu construire le modèle pour l'extraction des syntagmes nominaux dans des textes en langue portugaise, présenté à la fin de cette thèse. La réflexion sur l'interface construite par le biais de la

maquette nous a conduit à un Système de Recherche d'Information Assistée par Ordinateur, une nouvelle gamme de SRI.

Introduction

Dès l'invention des ordinateurs les hommes sont à la recherche d'une manière efficace de gérer, de stocker, de diffuser et de rechercher l'information. Plusieurs méthodes et techniques de gestion et de traitement d'information ont été développées tout au long de ces années. Si aujourd'hui nous sommes dans un haut niveau d'informatisation, c'est parce que les hommes ont été suffisamment compétents pour développer et maîtriser la technologie (soit celle des matériels, soit celle de la communication, soit celle de la construction des logiciels ou soit celle de la gestion et du traitement de l'information). Il est vrai que la technologie de matériels et de la communication s'est développée rapidement. Grâce à ce développement, il est possible aujourd'hui avoir une grande disponibilité de mémoire primaire et secondaire, branchée aux ordinateurs et à un coût très bas. Ce n'était pas imaginable il y a quinze ans. De même cette technologie, envisageant le confort des utilisateurs, nous offre une panoplie d'options en termes d'accessoires périphériques, soit pour l'entrée de données, soit pour la diffusion de l'information. La numérisation de données est partout.

Or, bien que nous assistions à cette évolution remarquable — dans les domaines de l'informatique, de la communication et de l'information — nous nous rendons compte qu'il faut encore beaucoup progresser, surtout dans le domaine de l'information documentaire, plus spécifiquement dans celui du traitement et de la dissémination de l'information textuelle. Plusieurs recherches sont en cours de développement dans ce domaine, et pourtant, les problèmes d'indexation automatique et de recherche d'information sont encore très actuels.

C'est là notre souci. Nous nous intéressons au domaine du traitement et de la dissémination de l'information. Le domaine de l'information est très vaste. Les problèmes que nous allons poser et étudier dans cette recherche ne concernent que la partie relative à l'information textuelle.

Nous allons, donc, étudier les problèmes de la recherche d'information en envisageant de proposer un nouveau modèle de système de recherche d'information (SRI). C'est-à-dire, un modèle complet de SRI pour des bases de données textuelles, plus couramment appelées bases de données texte plein (full text). On entend par modèle complet, ce qui est composé non seulement d'une interface de recherche d'information, mais aussi d'un module de traitement et d'indexation automatique d'information intégré à un module d'interface de recherche d'information.

La démarche suivie est reflétée dans ce document. Celui-ci est partagé en trois grandes parties. La première, appelée "La Problématique", est consacrée à la discussion des problèmes de la recherche d'information et à la définition d'un système de recherche d'information. Dans cette partie nous étudierons les problèmes concernant la faiblesse de précision des résultats d'une recherche d'information et le manque de convivialité de ces systèmes. En fait, ce premier chapitre essaye de montrer le contexte de cette recherche.

Dans le deuxième chapitre nous proposerons ce que nous appelons Système de Recherche d'Information Assistée par Ordinateur. Cette proposition est faite dans le but de résoudre les problèmes discutés dans le premier chapitre.

La deuxième partie est consacré à la construction d'une maquette d'un système de recherche d'information basée sur la proposition faite dans le deuxième chapitre.

Le troisième chapitre est consacré à la description de la procédure de construction d'une base de données indexée par syntagmes nominaux. Ce chapitre décrit les critères utilisés pour la constitution du corpus qui va intégrer cette base de données, les traitements préalables pour la mise en place de cette base de données et les remarques sur la procédure d'extraction des syntagmes nominaux. Nous avons fait cette extraction de manière artisanale, c'est-à-dire manuellement, étant donné l'inexistence d'un système d'extraction automatique des syntagmes nominaux dans des textes en langue portugaise.

Le quatrième chapitre décrit les procédures de construction de la maquette d'un système de recherche d'information basé sur l'utilisation des syntagmes nominaux comme moyen d'accès à l'information. Nous montrons dans ce chapitre la structure de données pour la base de données et la description des rapports entre les syntagmes nominaux dans leurs différents niveaux. C'est-à-dire la structure arborescente des syntagmes nominaux. Ceci montre la démarche de navigation dans ces structures.

Le cinquième chapitre est consacré à la mise en service de la maquette. C'est-à-dire la mise en place de la base de données à partir du corpus et l'indexation de cette base à l'aide de l'utilisation des syntagmes nominaux extraits du corpus choisi. Etant donné le fait que les syntagmes nominaux ont été extraits manuellement, l'indexation de la base a été également faite à la main. Nous profitons de ce chapitre pour faire aussi quelques remarques sur les problèmes rencontrés lors de cette procédure d'indexation et de même sur l'exploitation de la maquette.

Dans le sixième chapitre nous avons fait une expérimentation, en utilisant un thesaurus, en langue portugaise, appelé TCI – *Tesauros de Ciência da Informação* (Thesaurus de Science de l'Information), pour étudier le comportement de la maquette et aussi des syntagmes nominaux comme moyen d'accès à l'information.

La troisième et dernière partie de la thèse est vouée à la construction d'un modèle pour la reconnaissance et l'extraction automatique des syntagmes nominaux, dans les textes en langue portugaise. La présentation de ce modèle est précédée par une esquisse du nouveau SRI.

Ainsi, le septième chapitre, présente une étude d'un des problèmes rencontrés dans la procédure d'extraction manuelle de syntagmes nominaux : l'absence assez fréquente de déterminants dans les syntagmes nominaux en langue portugaise. Nous essayons de trouver des marques qui puissent aider à la procédure de reconnaissance et d'extraction automatique des syntagmes nominaux.

Le huitième chapitre est voué à l'esquisse du nouveau SRI, c'est-à-dire, le Système de Recherche d'Information Assistée par Ordinateur. Nous proposons, en grandes lignes, une démarche de développement de ce système, en donnant les indications de sa composition ainsi que le dessin des structures de données nécessaires pour soutenir l'organisation de syntagmes nominaux. Ce sont les indices qui permettront aux usagers naviguer dans la base de données et retrouver l'information.

Dans le neuvième chapitre nous établissons la première partie du modèle pour la reconnaissance et l'extraction automatique des syntagmes nominaux, en langue portugaise : la grammaire de référence. C'est-à-dire que nous avons construit une grammaire de référence qui fait la description ou plutôt la caractérisation de chaque unité lexicale. C'est la première partie du modèle de reconnaissance et d'extraction automatique des syntagmes nominaux. C'est là où nous faisons la description de caractéristiques de chaque unité lexicale, de chaque mot pour reconnaître les éléments d'un syntagme nominal.

Le dixième chapitre est consacré à l'établissement de la grammaire de reconnaissance et d'extraction automatique des syntagmes nominaux. Nous faisons, dans ce chapitre, la description de la méthodologie utilisée pour l'établissement des règles de réécriture des syntagmes nominaux, et la description de chaque règle de réécriture, lesquelles font partie de la grammaire.

La démarche présentée est calquée sur la progression chronologique du travail de recherche : nous sommes partis de la pratique et nous sommes arrivés à la fin en concluant avec une partie théorique. Ce parcours a été fondamental pour la recherche, car il a permis de connaître le comportement et la composition des syntagmes nominaux dans les textes en langue portugaise. Par ailleurs, il a permis aussi de comprendre le comportement des syntagmes nominaux dans un système de recherche d'information. Ainsi, le savoir-faire obtenu lors de la procédure d'extraction manuelle des syntagmes nominaux a été important pour la construction du modèle de reconnaissance et d'extraction automatique des syntagmes nominaux. Un exemple montre cette importance, celui d'avoir rencontré un pourcentage important de syntagmes nominaux sans déterminant dans le corpus. Un autre exemple c'est que nous avons pu nous rendre

compte de différences et de similitudes entre les langues française et portugaise, puisque nous connaissions d'abord la grammaire pour la reconnaissance et l'extraction automatique pour la langue française. Ces expériences et connaissances nous ont donc permis de concevoir le modèle de reconnaissance et d'extraction automatique des syntagmes nominaux pour la langue portugaise.

Nous avons mis dans l'annexe, le deuxième volume de cette thèse : a) les articles du corpus ; b) les syntagmes nominaux extraits du corpus d'articles ; c) le tableau avec les descriptions de SN consolidées ; d) un sous-ensemble de SN ajouté de la description de chaque SN. L'ensemble complet de description de SN est mis dans la disquette qui accompagne cette thèse.

Première Partie : la problématique

« Un prince sage donne aux choses les noms qui leur conviennent, et chaque chose doit être traitée d'après la signification du nom qu'il lui donne. »
Confucius, Entretiens, VII, 13.

Chapitre Premier Systèmes de Recherche d'Information : contexte de la recherche

Depuis que les ordinateurs sont apparus, des milliards d'informations y ont été enregistrées dans plusieurs bases de données, dans divers domaines de connaissances et sous diverses formes (numériques, textes, images etc.). Etant donné que les ressources informationnelles sont de plus en plus accessibles aux utilisateurs personnels, le principal problème aujourd'hui est de savoir comment accéder à l'information dont on a besoin.

Plus récemment, on a vu l'apparition des Nouvelles Technologies de l'Information et de la Communication (NTIC), poussées par les réseaux des réseaux, Internet, ainsi que par le programme américain *National Information Infrastructure* (NII). Le résultat de cette initiative est la croissance phénoménale du volume d'informations aussi bien que des outils de traitement et de recherche d'information.

D'après Pierre LÉVY : ***« La prospérité des nations, des régions, des entreprises et des individus dépend de leur capacité à naviguer sur l'espace du savoir. La puissance est désormais conférée par la gestion optimale des connaissances, qu'elles soient techniques, scientifiques, de l'ordre de la communication ou qu'elles relèvent de la relation "éthique" avec l'autre. »*** [1] .

Les contextes présentés sont, en fait, les mêmes. En premier lieu on observe trois phénomènes : la baisse constante du marché du matériel informatique, l'insertion des NTIC et la production croissante d'informations. Par ailleurs, on insiste actuellement sur l'importance de la connaissance, du savoir pour les nations, les régions, les peuples, c'est-à-dire l'utilisation de l'information est aujourd'hui au cœur des préoccupations. Selon Pierre LÉVY, l'arrivée des NTIC est à l'origine d'un changement de paradigmes. La gestion optimale des connaissances passe nécessairement par l'acquisition et la production de l'information. C'est là qu'on voit l'importance des Systèmes de Recherche de l'Information (SRI), car ce sont des outils qu'on utilise pour la recherche d'informations.

Nous verrons tout au long de ce chapitre la description des caractéristiques de ces systèmes, comment ils fonctionnent et quels problèmes ils entraînent pour les usagers. Tout d'abord, nous définirons ce qu'est un SRI en montrant quelques définitions données par des spécialistes en SRI, puis nous situerons les SRI par rapport aux Systèmes d'Information, ensuite nous discuterons les problèmes liés à l'utilisation de ce type de système, enfin nous présenterons une typologie de ces systèmes avant de conclure.

1 Définitions d'un SRI

Il y a plusieurs définitions d'un SRI, lesquelles se ressemblent plus ou moins.

Harter définit un SRI comme ***"un dispositif qui s'interpose entre les usagers potentiels et la collection d'informations"*** [2] . De plus, Tomek Strzalkowski dit que ***« la tâche typique de la recherche d'information, c'est la sélection des documents dans une base de données, en réponse à une requête de l'usager, et leur rangement par ordre de pertinence »*** [3] . Tandis que Alan Smeaton donne la définition suivante : ***« le but d'un système de recherche d'information est de récupérer des documents en réponse à une requête des usagers, de manière que les contenus des documents répondent pertinemment au besoin originel d'information de l'usager »*** [4] . Or, ces définitions ressemblent plutôt à la définition d'une Interface [5] de Recherche d'Information (IRI), car elles n'explicitent pas les procédures de traitement de l'information et

[1] Pierre LEVY. *L'Intelligence collective : Pour une anthropologie du cyberspace* . Paris : La Découverte, 1997. p. 17.

[2] « An *information retrieval system* is a device interposed between a potential user of information and the information colection itself. » Stephen P. HARTER. *Online Information Retrieval: Concepts, Principles and Techniques* . Orlando : Academic Press. Inc., 1986. p. 2.

[3] « A typical information retrieval (IR) task is to select documents from a database in reponse to a user's query, and rank these documents according to relevance. Tomek STRZALKOWSKI. " Natural language processing in large-scale text retrieval tasks ". In. : *Text REtrieval Conference (TREC-1).* Gaithersburg, 1993. p. 173.

d'indexation automatique des documents. Ce sont des définitions qui prennent en compte ce que les usagers perçoivent d'un SRI, Tomek STRZALKOWSKI et Alan SMEATON explicitant montre ce qu'un SRI doit leur offrir. Or, il y a tout un ensemble de procédures pour que ces usagers puissent accéder à l'information. Ainsi, un SRI est plus qu'une simple IRI.

Selon SALTON & MCGILL, ***« un SRI traite de la représentation, du stockage, de l'organisation et de l'accès aux items d'information »*** [6] . Malgré sa simplicité, cette définition est plus complète, car elle caractérise aussi bien l'interface de recherche d'information que les aspects concernant le traitement et le stockage de l'information.

On peut dire alors qu'un SRI est plutôt un système composé d'une part par un module chargé du traitement, de l'indexation et du stockage de l'information. Ce module construit, à partir du traitement de l'information, une structure de données organisées de manière à permettre l'accès rapide à l'information. D'autre part, il est composé par un module, aussi appelé interface, qui sert à interagir avec les usagers, dotée des mécanismes de sélection d'information orientés par les requêtes formulées par les usagers. On distinguera, donc, ce qu'est un SRI et ce qu'est un IRI.

2 SRI et les Systèmes d'Information

Il y a plusieurs types de systèmes d'information sur le marché, lesquels sont classés par Gerald SALTON & Michael J. MCGILL en cinq grands groupes : les systèmes de recherche d'information, les systèmes de gestion d'information, les systèmes de gestion de bases de données, les systèmes de support à la prise de décision, les systèmes de question-réponse. Pour eux, les systèmes de gestion d'information sont une sorte de système de gestion de bases de données appropriées aux administrateurs, enrichis de graphiques, analyses et synthèses.

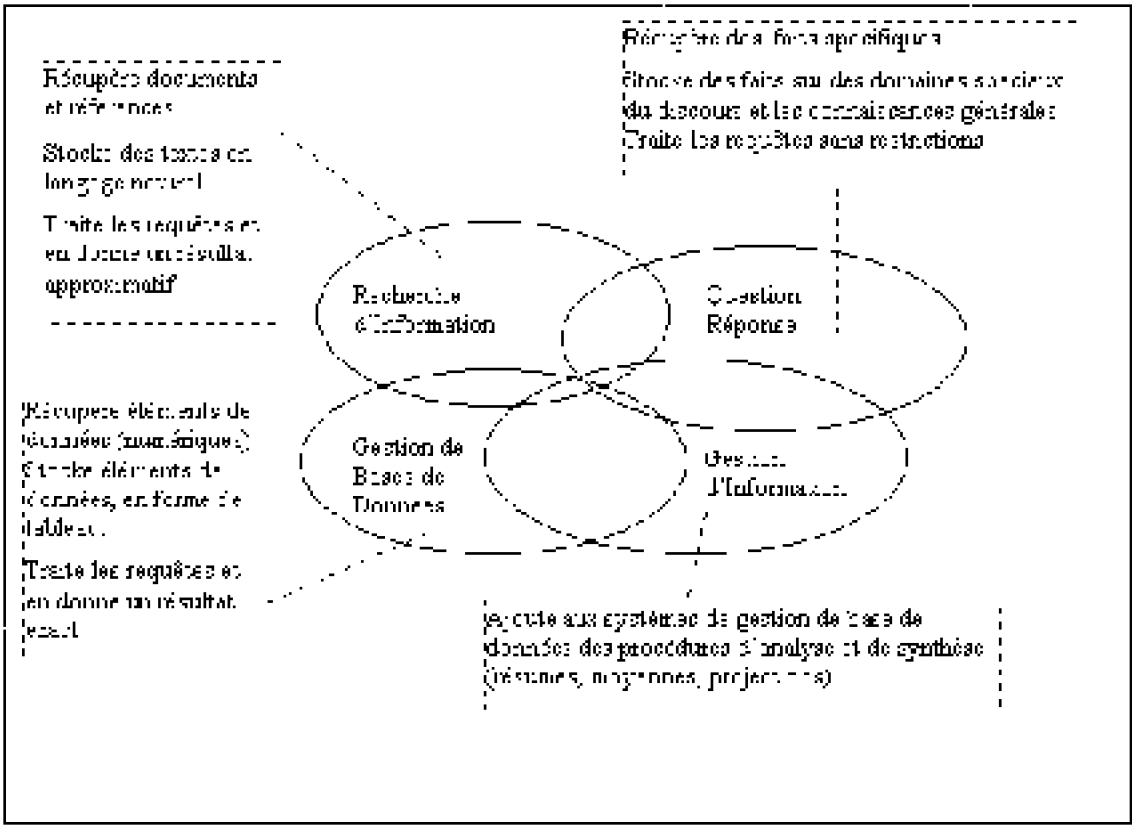
Selon Gerald Salton & Michael J. McGill, les systèmes d'information ***« …exécutent des opérations spécifiques sur des classes homogènes d'items d'information. Normalement les SRI ne remplissent pas la fonction de gestion d'information et vice-versa. Cependant, en principe, on peut concevoir aussi une autre sorte de système, en rassemblant une variété de composants différents, dans une simple structure coopérative, incluant le SRI, le SGBD, systèmes de graphiques et d'autres outils techniques. La conjonction de ces systèmes en fait rend un outil puissant***

[4] « The aim of an information retrieval system is to retrieve documents in response to a users request in such a way that the content of the documents will be relevant to the user's original information need. ». Alan F. Smeaton. "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, 21-23 Mars, 1989, p. 2.

[5] On entend le mot interface comme le moyen de communication entre l'usager et l'ordinateur.

[6] « Information retrieval (IR) is concerned with the representation, storage, organization and accessing of information items. ». Gerald SALTON & Michael J. McGILL. *Introduction to modern information retrieval*. New York : Mcgraw-Hill Book Company, 1983, p. 1

pour la prise de décision, ce sont les systèmes de support à la prise de décision » [7] . Ils ont construit le schéma [8] de la figure 1.1 pour montrer la similarité entre ces systèmes.



Les évolutions récentes des technologies de l'information et de la communication ont changé un peu ce panorama (le schéma a été extrait d'un livre sorti en 1983). Divers types de systèmes sont apparus récemment, mais dans l'essence, en ce qui concerne

[7] « …perform specific operation on homogeneous classes of information items. Normally, information retrieval systems do not perform management information functions, and vice versa. However, it is in principle possible to conceive of information systems in which a variety of different components are assembled into a single cooperating structure that includes information retrieval systems, data base management systems, computer graphiques systems, and other technical capabilities which collectively provide powerful tools in support of the decision making process. ». Gerald SALTON & Michael J. McGILL. *Introduction to modern information retrieval*. New York : Mcgraw-Hill Book Company, 1983, p. 9

[8] Gerald SALTON & Michael J. McGILL. *Introduction to modern information retrieval*… 1983, p. 10.

l'information, ils peuvent être représentés par un des systèmes présents dans la figure 1.1. Par exemple, les systèmes hypertextes et hypermédia peuvent être encadrés dans les systèmes de recherche d'information, car leur fonction primordiale est la dissémination de l'information. Il faut peut-être adapter la définition des SRI, en ajoutant l'image et le son comme type d'information traité. Le but de présenter ce schéma est plutôt de montrer la similarité des SRI avec les systèmes d'information d'une manière générale, et non de présenter exhaustivement tous les systèmes d'information et de les définir de manière approfondie.

Il y a quelques différences entre les SRI et les systèmes de gestion de bases de données (SGBD). Parmi celles-là on distingue :

- Les SGBD stockent des informations factuelles, normalement numériques. Tandis que les SRI stockent des textes intégraux (en langage naturel), des références bibliographiques (notices bibliographiques), des informations imagées, sonores (hypermédias) ;
- Les bases de données gérées par les SGBD ont une structure de données, où chaque enregistrement a des attributs avec des formats et valeurs précises. Dans les SRI, les bases de données ont une organisation différente à cause de la caractéristique des données qui y sont traitées et enregistrées. On peut dire qu'il y a un format général mais variable, soit en fonction de la taille de chaque champ, soit en fonction de l'absence ou présence d'un champ. Dans les cas de bases de données bibliographiques, des champs peuvent se répéter (ex.: auteur), ou peuvent apparaître dans une notice (registre) et être absents dans une autre notice (ex.: résumé). Les champs étant textuels, ils ont une taille variable ;
- La réponse donnée par les SGBD correspond exactement à ce qui l'utilisateur a demandé, puisque la recherche est faite sur les attributs d'un registre. Ainsi, par exemple : si on demande une recherche de tous les employés d'une entreprise qui gagnent plus de 20.000 francs, le SGBD fera la recherche sur l'attribut salaire du fichier d'employés dans la base de données. Le système donnera une réponse exacte, il ne donnera que les enregistrements des employés qui gagnent plus de 20.000 francs. Dans les SRI le résultat n'est pas aussi précis que celui donné par les SGBD. La réponse peut avoir des documents ou des références qui ne correspondent pas au besoin d'information de l'utilisateur. Malgré le traitement exact de la requête, le résultat ne l'est pas, il est approximatif. Ainsi, la différence entre les résultats d'un SRI et ceux d'un SGBD vient de ce que dans ces derniers les attributs ont des valeurs bien précises. Ce qui n'arrive pas aux bases de données textuelles, c'est le fait qu'un champ textuel n'a pas une valeur unique aussi bien définie qu'un champ numérique et que la recherche est faite dans des textes à travers une composition de mots avec des opérateurs booléens. Il est vrai que dans une base de données bibliographiques nous avons une structure qui nous permet de faire une recherche précise sur le nom de l'auteur, sur le nom de l'éditeur, utilisant bien sûr, une requête avec le nom complet, soit de l'auteur ou de l'éditeur. Cependant, on ne peut pas le faire de la même façon sur un champ comme le résumé. La taille d'un résumé est variable et elle peut arriver à des centaines de mots. Ce qui rend difficile de mettre

tout le champ résumé dans une requête. C'est une des raisons pour lesquelles on utilise une requête composée de mots et d'opérateurs booléens. Pour mieux illustrer le problème de l'inexactitude d'une réponse des SRI, on donnera les exemples suivants :

– Dans la base de données de publications périodiques Myriade, lorsqu'on demande une recherche dans le titre, pour trouver une revue appelée SCIENCE, en utilisant seulement le champ de titre, avec le mot SCIENCE dans la requête, le SRI donne comme réponse 2248 références. Parmi ces références, il y a 5 revues appelées SCIENCE. Tandis que dans les 2243 références restantes, le mot SCIENCE est présent dans le titre, mais ajouté d'autres mots. On est donc obligé de faire une recherche multicritère [9] pour trouver la revue cherchée ;
– Si on veut trouver des documents portant sur l'ANALYSE DE SYSTEMES D'INFORMATION dans une base de données texte plein, normalement on n'utilise que les mots *analyse*, *systèmes* et *information*. Or, le SRI trouvera aussi bien les documents parlant de l'Analyse de Systèmes d'Information que les documents ayant comme sujet les Systèmes d'Analyse d'Information.

· Ces deux exemples montrent ce qui nous avons parlé plus haut, c'est-à-dire, ils montrent dans la pratique que le résultat d'une consultation dans un SRI n'est pas exact mais approximatif.

En ce qui concerne les systèmes de question / réponse, il s'agit de systèmes qui stockent des informations factuelles sur un domaine spécifique et ainsi que des connaissances et des faits s'y rapportant. Normalement ce type de système peut accepter des requêtes en langage naturel et éventuellement donner aussi la réponse, à cette requête, en langage naturel. En fait, ce type de système, prend la requête, l'analyse et la compare en utilisant des connaissances et des faits stockés.

On voit, ainsi, que les points de similitudes entre tous ces systèmes sont les suivants :

· Ils ont derrière eux des structures de bases de données ;
· Ils ont une fonction spécifique de traitement et de stockage de données ;
· Ils ont une fonction de recherche d'information spécifique à chaque type de système ou d'application. En conséquence, ils ont une interface de recherche d'information, soit guidée par un menu, soit par un langage. Ce qui ressemble aux SRI.

On entend comme base de données, un ensemble de données, organisées en structures de données capables de faciliter la mise à jour, le traitement et principalement la diffusion de l'information.

3 Systèmes de Recherche d'Information : caractéristiques et faiblesses

[9] Une recherche multicritére est basée sur l'utilisation de requêtes avec plus d'un champ de recherche. S'on utilise dans une requête le champ de résumé et le champ d'éditeur pour trouver une information donnée, on a là une recherche multicritère.

On trouve dans la littérature spécialisée plusieurs références traitant des Systèmes de Recherche d'Information.

Les SRI traitent et diffusent des informations du type textuelles, images et sonores. Comme information textuelle on entend les informations référentielles et les informations primaires. Les informations référentielles sont celles qui font référence à un objet donné. En ce sens, un objet peut être un livre, un article, un document, une personne, une entreprise, une entité ou une autre chose. Parmi les bases de données qui stockent ce type d'information (référentielle), on trouve : des bases de données bibliographiques (notices bibliographiques), des répertoires d'entreprise, des bases de données d'outils, des bases de données de personnes ou des bases de données de logiciels. C'est-à-dire que ces bases de données ne contiennent pas l'objet en question, mais des informations concernant ces objets comme : titre, auteur, éditeur, résumé, lieu de publication (pour les notices bibliographiques).

Les informations primaires sont celles qui sont à l'intérieur des objets eux-mêmes, celles qui font partie de ces objets. Ici, on parle plutôt des textes qui appartiennent à un livre, à un document, à un article, à un exemplaire d'un journal. Comme exemple de bases de données qui contiennent des informations primaires, on a les bases de données texte plein comme, dans la base de donnée du journal LE MONDE, où on trouve toutes les nouvelles publiées à une période donnée.

Ainsi, dans ce travail, nous ne traitons que des SRI qui font la recherche appliquée aux bases de données texte plein. Ces bases de données sont celles qui contiennent des textes intégraux comme des articles ou même des chapitres d'un livre, parmi d'autres types de documents textuels. Bien qu'on ne traite que du texte, ces documents peuvent être illustrés avec des images et des sons.

Les SRI sont devenus de plus en plus communs — les moteurs de recherche sont un des exemples de SRI les plus répandus — ainsi que les problèmes que ces systèmes entraînent auprès des utilisateurs. Les utilisateurs voudraient toujours travailler avec un SRI convivial et qui leur donne des réponses précises. Et pourtant, ce n'est pas ce qu'ils y trouvent. Ce sont justement les deux plus grands problèmes que les SRI posent aux utilisateurs, le manque de convivialité des interfaces de recherche d'information et les résultats peu satisfaisants donnés par les SRI. Ce sont les raisons qui amènent, aujourd'hui, beaucoup de chercheurs à travailler sur la construction de méthodes d'évaluation des SRI et de l'évaluation elle-même. Et aussi dans le développement de nouveaux SRI, en prenant en compte des nouvelles approches d'indexation et de dessin d'interface homme-machine.

Afin de mieux connaître ces systèmes on va étudier d'abord leurs composants et on discutera, ensuite pour chaque composant principal, des problèmes de son utilisation.

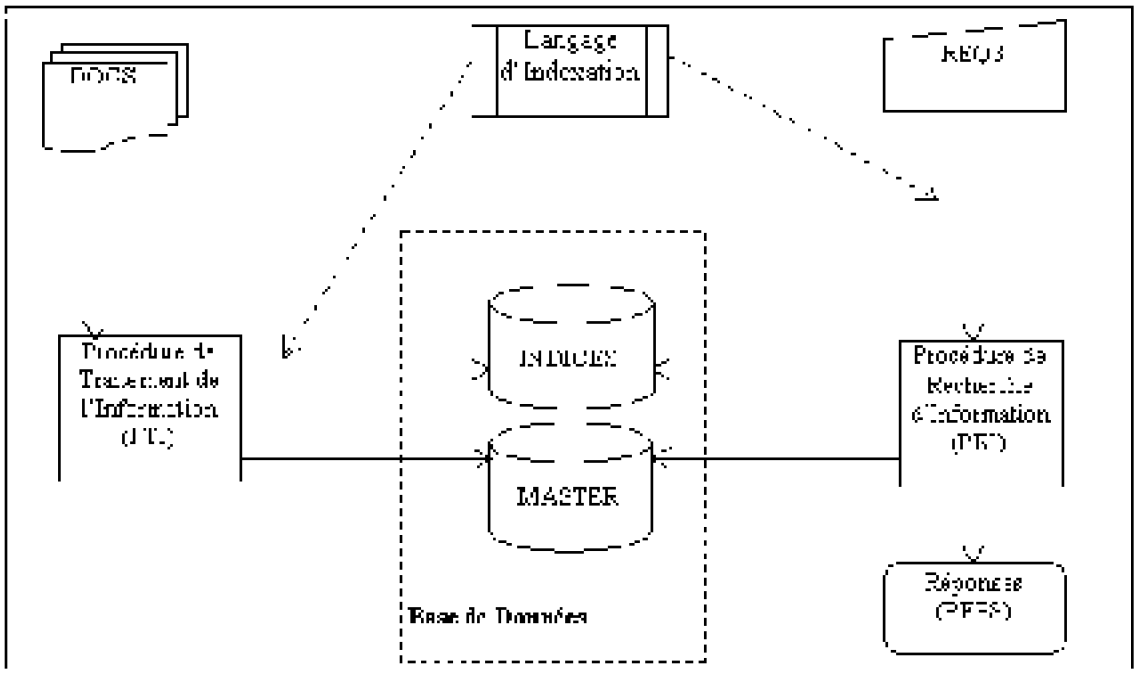
3.1 Composants d'un SRI

Les composants fonctionnels d'un SRI, pour distinguer ses principales fonctions, peuvent se regrouper en des modules majeurs. On peut les décrire comme un système constitué : a) d'un ensemble d'information (DOCS) ; b) d'une Procédure de Traitement d'Information (PTI) ; c) d'un ensemble de requêtes (REQS) ; et d) d'un mécanisme qui détermine

quelles informations répondent aux requêtes — Procédure de Recherche d'Information (PRI). Ce qui peut être montré par le schéma de la figure 1.2.

La PTI est la procédure responsable par le traitement, par l'indexation et par le stockage des documents. Comme résultat elle génère deux structures de données ou deux groupes de fichiers : 1) INDICES, où sont stockés les descripteurs dans une structure construite de manière à faciliter l'accès à l'information ; 2) MASTER où les contenus de documents sont stockés. L'extraction des descripteurs est faite en ayant comme base un Langage d'Indexation. Ce langage peut être soit pré-spécifié (vocabulaire contrôlé), soit pris librement dans les documents de la base de données (vocabulaire non contrôlé ou termes libres).

La PRI est la procédure chargée de recevoir la requête de l'usager (REQS), de l'interpréter, de déterminer la similitude des items d'information selon ce que demandent ces requêtes, et de donner les réponses à l'usager (REPS). Il faut que la requête soit composée de termes utilisés dans le langage d'indexation et pour l'indexation de la base de données, sinon on échoue complètement dans la recherche. On voit là que l'Interface de Recherche d'Information fait partie de la PRI.



Il faut remarquer que dans la base de données (BD) on peut avoir d'autres fichiers, ceux qu'on appelle habituellement de fichiers auxiliaires, pour simplifier le schéma, ils ne sont pas représentés dans la figure. Une autre remarque concerne la séquence d'exécution de ces deux procédures. D'abord, il faut que les documents de la base de données soient traités et structurés pour que les utilisateurs puissent faire la recherche d'information. La PTI est donc, naturellement, la première procédure à être exécutée. Les deux procédures doivent être exécutées dans des moments différents car il faut que les structures d'accès à l'information soient prêtes et cohérentes pour que la PRI puisse faire la recherche. La mise à jour de données faite en même temps qu'une recherche peut entraîner de mauvais résultats, c'est pourquoi, pour maintenir la cohérence de ceux-ci, il

faut que la mise à jour de données soit faite, en dehors et avant que les usagers les utilisent.

3.2 Procédure de traitement de l'information (PTI)

Selon ce qu'on a vu dans le schéma de la figure 1.2 et de la description, les PTI sont à la base de la préparation et de l'indexation des documents d'une base de données. On dit que le résultat de l'indexation est la *représentation du contenu* des documents indexés.

A quoi sert la *représentation du contenu* des documents d'une base de données ou les indices produits par l'indexation ? Bien que le terme *représentation du contenu* soit utilisé couramment dans les cours de sciences de l'information et de la communication — et ainsi que dans la littérature sur l'indexation [10], [11] et [12], — ceci pour désigner le résultat de l'indexation automatique, M. Le Guern rappelle que l'utilisation de ce terme est inexacte. La *représentation du contenu* d'un document est le texte lui-même tandis que le résultat d'une indexation se présente comme des mots ou des parties du texte du document en fonction du type d'indexation.

Certains SRI font la recherche d'information directement dans les textes pleins. Cela peut bien fonctionner pour une petite base de données, mais dès que le nombre de documents de la base de données commence à s'agrandir, cette méthode deviendra peu efficace, car le temps de réponse va sûrement augmenter. Or, la raison principale pour laquelle on indexe les documents d'une base de données c'est pour réduire le plus possible le temps de réponse d'une recherche d'information. C'est à partir de ces indices qu'un SRI peut trouver des informations en réponse à une requête. Le résultat de l'indexation n'est donc pas la représentation des contenus des documents, mais des « pistes » extraits des documents de manière à permettre aux usagers de les retrouver postérieurement.

La performance, en tant que temps de réponse d'un SRI est lié à la structuration du stockage des indices avec les adresses des documents d'où chaque indice a été extrait. Par ailleurs, la performance du SRI, en tant que précision de la réponse, est liée principalement à la technique et à la méthode d'indexation.

Comme technique d'indexation on entend la manière de le faire. On distingue donc

[10] « L'indexation se définit comme l'activité consistant à représenter le contenu d'un document… ». Georges VAN SLYPE. *Les langages d'indexation : conception, construction et utilisation dans les systèmes documentaires.* Paris : Les éditions d'Organisation, 1987. p. 21.

[11] « In extracting indexing, words or phrases appearing in a text are extracted and used to represent the content of the text as a whole. ». Frederic W. Lancaster. *Indexing and Abstracting in Theory and Practice.* London : Library Association Publishing Ltd., 1991. p. 221.

[12] « Il est souvent dit de l'indexation qu'elle donne une représentation du contenu d'un document ». Richard BOUCHÉ, Sylvie LAINÉ & Jean-Paul METZGER. « 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, p. 4.

trois techniques d'indexation :

1. indexation manuelle ou indexation faite par des personnes Cette technique d'indexation est souvent utilisée pour l'indexation des bases de données bibliographiques ou bases de données référentielles. C'est une technique dépendante des personnes, car sont eux qui font l'indexation, à partir de l'analyse des documents. De cette analyse sortent les descripteurs, soit attribués à l'aide d'un vocabulaire contrôlé (thesaurus et d'autres langages d'indexation), soit attribués librement. Ceux qui s'occupent de ce type d'indexation sont des documentalistes ou souvent des spécialistes du domaine de connaissance de la base de données ;
2. indexation mixte ou indexation assistée par ordinateur A l'instar de la technique d'indexation manuelle, cette méthode est aussi souvent utilisé pour l'indexation des bases de données bibliographiques. Tout d'abord l'ordinateur fait l'indexation des documents existant dans la base de données et après un documentaliste ou spécialiste du domaine de la base de données revoit l'indexation la complétant et/ou la corrigeant. Là encore on utilise des vocabulaires contrôlés pour aider l'indexation ;
3. indexation automatique C'est une technique d'indexation qui peut être utilisé autant pour des bases de données bibliographiques que pour des bases de données texte plein. Du point de vue technique, rien n'empêche de faire l'indexation des bases de données textes pleins par les deux techniques précédents. Le problème principal est le temps de traitement d'un document texte plein. Il est plus économique et plus rapide de les indexer par des techniques automatiques. Ceci est justifié, non seulement par la taille de ces documents mais plutôt par leur volume, lesquels ont augmenté, aujourd'hui de manière phénoménale. C'est réalisable aussi à l'aide de vocabulaires contrôlés.

On entend pour méthode d'indexation, la manière d'extraire et de traiter les indices. Nous allons présenter deux méthodes d'indexation automatiques de documents pour la recherche d'information textuelle. En fait, il en a plusieurs, qui ne sont que des variations de ces deux méthodes. Nous traiterons seulement ici de l'indexation automatique pour la recherche d'information textuelle primaire [13] .

3.2.1 Indexation automatique par mots clés

C'est la méthode traditionnelle d'indexation automatique d'une base de données textuelle, soit de notices bibliographiques, soit de textes pleins. D'une manière générale, cette méthode consiste à extraire simplement les mots existants dans un document de la base de données. Ceux-ci sont aussi appelés mots clés ou descripteurs. La variation qu'on trouve dans cette méthode concerne le traitement de ces mots après leur extraction. Ainsi, on distingue deux de ces méthodes :

1. les méthodes élémentaires comme la simple extraction des mots où ces mots sont des descripteurs en excluant les mots vides ;

[13] L'information textuelle primaire comprend les textes d'un article, d'un livre, ou d'un autre type de documents textuels.

2. les méthodes statistiques qui consistent à définir un modèle probabiliste des occurrences des mots dans un document considéré à l'intérieur d'une collection bien définie de manière à établir leur caractère pertinent pour participer à la description du contenu.

La façon dont les descripteurs sont définis et structurés induit d'une certaine manière l'utilisation d'un type d'interface de recherche d'information spécifique. Ainsi, si les descripteurs sont organisés de manière hiérarchique, on peut utiliser une interface guidée par des menus ou même une autre guidée par un langage de commande. Par contre, si les descripteurs sont composés de mots simples, sans hiérarchie les uns par rapport aux autres, l'utilisation d'une interface guidée par des menus n'est certainement pas la plus appropriée. On aura une quantité importante de mots qui n'offrent pas une organisation adéquate pour composer les menus, sauf leur organisation par ordre alphabétique. Cela ressemble à une recherche de mots dans un dictionnaire.

Une recherche utilisant une liste de mots demande une interface capable d'offrir aux utilisateurs la possibilité de combiner les mots les uns par rapport aux autres, pour exprimer le besoin d'information de l'utilisateur. Ainsi, les SRI traditionnels les plus répandus sont ceux dont l'interface est guidée par un langage de commandes où les requêtes sont formulées à l'aide des expressions composées par des mots, par des opérateurs booléens [14] , et par d'autres opérateurs comme ceux de troncatures [15] et de proximité [16] . Or, selon la littérature spécialisée et dans la pratique même, on se rend compte que cette approche entraîne des problèmes de précision des résultats donnés par les SRI sur une base de données textuelles.

Certains SRI permettent l'utilisation d'opérateurs de proximité ou même l'insertion des poids (déterminés automatiquement ou par définition de l'usager) sur les mots clés. Or, les mots ont des caractéristiques qui empêchent l'obtention de bons résultats dans la procédure de recherche d'information. Le premier aspect c'est qu'un mot peut avoir des signifiés différents, selon le domaine. Par exemple : *goutte* peut indiquer une très petite quantité de liquide qui prend une forme arrondie ; alors que dans le champ de la médecine il indique une sorte d'inflammation douloureuse des articulations. Un autre aspect c'est que les mêmes mots peuvent être utilisés dans des phrases différentes, tout en ayant des liaisons différentes et exprimer des concepts totalement dissemblables.

[14] Opérateur booléen sont des opérateurs utilisés dans la logique booléenne comme ET, OU ou SAUF. L'opérateur ET fait l'intersection entre deux ensembles. L'opérateur OU fait l'union de deux ensembles et l'opérateur SAUF fait l'exclusion d'un ensemble de l'autre.

[15] L'opérateur de troncature permet la spécification d'un masque dans une requête. C'est-à-dire, cet opérateur permet de faire une recherche à partir d'un préfixe ou d'un suffixe d'un mot. Exemple : inf+ Ici nous avons un opérateur de troncature (+) à droite. Cela indique au SRI que l'utilisateur veut tous les documents dont les descripteurs ont le préfixe INF. Ainsi, le SRI pourra trouver des descripteurs comme : INFORMATION, INFORMATIQUE, INFORMATIONNELLE, etc.

[16] L'opérateur de voisinage permet aux utilisateurs de construire des requêtes indiquant la position relative des mots, les uns par rapport aux autres dans les textes. Exemple : *information (w) scientifique*indique au SRI qu'il doit chercher les documents qui contiennent ces deux mots dans cette séquence.

Exemples: *Le traitement linguistique de l'information et le traitement de l'information linguistique* ; *l'analyse statistique de l'information et l'analyse de l'information statistique*. Ces deux exemples utilisent les mêmes mots, dans chaque phrase, mais dans un autre ordre ce qui entraîne des significations différentes. Un dernier aspect concerne le phénomène de la synonymie : des mots complètement différents peuvent être utilisés pour exprimer le même concept. Par exemple, les termes *tremblement de terre* et *séisme* ont le même sens mais les mots utilisés sont complètement distincts.

Selon SMEATON [17] , ***« Les approches conventionnelles utilisées en Recherche d'Information telles que l'indexation basée sur les mots et la recherche par expressions booléennes [18] ne peuvent pour autant pas résoudre ce type de problème »*** . Il semble que dans la mesure où les problèmes de synonymie et de polysémie de mots existent toujours, il nous semble inutile de tenter de les résoudre. Le but de la recherche d'information est plutôt de trouver l'information dont l'utilisateur a besoin ou qu'il désire.

3.2.2 Indexation par unités complexes [19]

Cette méthode d'indexation privilégie la phrase, et non les mots pris isolément, dans la procédure d'indexation et de recherche d'information. Son objectif est d'obtenir, à partir de l'analyse linguistique des documents, les structures syntaxiques qui sont extraites et organisées sous forme d'arbre ajouté aux informations syntaxiques des unités de la phrase. Cette structure est utilisée dans la procédure de détermination de la similarité entre les documents et les requêtes. Comme exemple d'utilisation de cette approche, on trouve le projet ESPRIT SIMPR [20] .

Dans un autre côté, selon BOUCHÉ, l'approche développée au sein du groupe SYDO part du principe ***que « le lexique, en tant que composant de la langue, ne contient que des éléments qui sont des propriétés, c'est-à-dire des prédicats. Le mot est***

[17] Alan SMEATON. " Prospects for intelligent, languaged-based information retrieval ". *Online Review*. 1991, vol. 15, n°. 6, p. 374.

[18] ***Expressions booléennes : dans ce contexte, c'est une expression composée par des descripteurs et des opérateurs booléens (OU, ET ou SAUF). Normalement une requête peut être composée par des opérateurs booléens ainsi que des opérateurs de proximité.***

[19] **Unités complexes dans ce contexte c'est des unités d'information représentées par des phrases ou des syntagmes nominaux.**

[20] Ce projet utilise la phrase comme moyens d'accès à l'information. C'est-à-dire qu'il fait l'indexation et la recherche d'information par le biais d'utilisation des phrases. Ces phrases sont organisées dans une structure que s'appelle TSA - *Tree Structured Analitics* (arbres analytiques structurés). Cette structure consiste en un arbre binaire où sont stockées des informations comme : mots originels de la phrase, leurs formes de base, étiquettes de catégorie lexique, morphologique et syntaxique. La procédure de recherche d'information utilise cette structure pour trouver les informations demandées par les requêtes. Ces requêtes sont constituées de phrases. C'est-à-dire, l'unité utilisée dans une requête est une phrase et non pas les mots comme dans les SRI classiques. SMEATON, Alan F. et SHERIDAN, Paraic. " Using Morpho-Syntaxique Language Analysis in Phrase Matching ". *RIAO 9 : Recherche d'Information Assistée par Ordinateur*. Barcelona, 1991. vol. 1, p. 414-430.

donc un prédicat et il ne peut pas être considéré de façon isolée car il fait référence à un objet de la réalité extra-linguistique de l'auteur du document. Il ne peut pas exprimer 'ce dont parle le document'. Il ne peut donc pas être un descripteur » [21] .

LE GUERN étant le précurseur de l'approche développée par le groupe SYDO, corrobore cette approche, disant : ***« La finalité du descripteur exclut qu'on puisse l'envisager en faisant abstraction de la valeur référentielle de ses occurrences dans le corpus. Les mots de la langue, en tant qu'ils sont mots de la langue, ne signifient que des propriétés, jamais des entités ; ils signifient des attributs et non des substances, tant qu'ils ne sont pas mis en œuvre dans le discours. Le descripteur, quant à lui, signifie une entité, une substance au sens de la philosophie d'Aristote. Le descripteur ne peut donc pas être considéré, à l'instar des mots de la langue comme un symbole sans référence. »*** [22]

Le fait que le mot, pris isolément, soit un signe sans référence de même que l'idée que les descripteurs devraient être un signe avec références renforce la validité de l'approche menée par SMEATON, à savoir l'utilisation de la phrase comme descripteur au lieu d'utiliser les mots isolés.

Dans l'approche de l'utilisation d'une phrase comme descripteur, cela veut dire que les mots sont laissés dans leurs contextes, comme les auteurs les ont rassemblés, en opposition à l'indexation traditionnelle, où les mots sont détachés de leurs contextes.

Par analogie, l'approche adoptée par SMEATON ressemble à celle du groupe SYDO. Selon Richard BOUCHÉ, ***« … la plus petite unité du discours porteuse d'une valeur référentielle est le syntagme nominal. C'est elle qu'il importe d'identifier dans le document »*** [23] . C'est-à-dire que le descripteur doit être représenté par le syntagme nominal.

Quelques travaux ont été réalisés pour l'extraction des syntagmes nominaux dans des textes en langue française. Parmi eux nous citons les thèses de : J-P. METZGER, de Omar LAROUK, Marcilio DE BRITO.

3.3 Procédure de Recherche d'Information (PRI)

La PRI joue un rôle important dans un SRI car elle est chargée de faire l'intermédiation entre l'usager et la base de données. La bonne utilisation d'une base de données dépend de la convivialité de l'interaction entre le SRI et l'usager, ainsi que de l'efficacité de traitement de la requête fourni par l'usager. D'après ce qu'on a montré dans la figure 1.2,

[21] Richard BOUCHÉ. « Le Syntagme Nominal, une Nouvelle Approche des Bases de Données Textuelles ». *Meta*. 1989, vol. 34, n°. 3. p. 429.

[22] Michel LE GUERN. « 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 "*. Condensé, supplément I,Université Clermont Ferrand, 1982. p.165-166.

[23] Richard BOUCHÉ. " Le Syntagme Nominal, une Nouvelle Approche des Bases de Données Textuelles ". *Meta*. 1989, vol. 34, n°. 3. p. 430.

l'interface de recherche d'information (IRI) est la partie d'un SRI chargée de recevoir une requête, de la traiter, de déterminer la similitude entre la demande d'information contenue dans la requête et les items d'information de la base de données.

Jusqu'à récemment, avant l'insertion des NTIC, l'utilisateur final accédait à l'information au moyen d'un intermédiaire, un documentaliste, bibliothécaire ou technicien d'information. Ce modèle a été conçu comme tel, d'une part, parce qu'on avait quelques difficultés de télécommunication pour se connecter aux systèmes de banques de données. Il n'y avait pas de disponibilités des réseaux d'ordinateurs comme on les trouve aujourd'hui. D'autre part, on avait du mal à accéder aux banques de données, car il fallait maîtriser un grand nombre de connaissances. Parmi toutes les commandes existantes il fallait maîtriser celles qui connectaient au système hôte hébergeant la banque de donnée ; celles qui mettaient en opération le système de recherche d'information ; celles ouvrant une session de recherche ; celles construisant une requête ; celles qui montraient les documents ou références résultant de cette requête ; celles qui les imprimaient. De plus, il fallait connaître la logique booléenne et ses opérateurs ; le langage d'indexation (thésaurus, vocabulaire contrôlé) ; les noms des champs de recherche. Enfin, la grande complexité de ses outils exigeait l'intermédiation d'une personne spécialisée dans le domaine de la recherche d'information pour qu'on puisse accéder à l'information.

Aujourd'hui les aspects essentiels concernant les commandes qui mettent un ordinateur en opération ou même qui connectent aux réseaux ne posent plus de problèmes aux utilisateurs, parce que l'ordinateur individuel est maintenant très répandu et les réseaux d'ordinateur de plus en plus disponibles au grand public. Les logiciels de connexion aux réseaux et les systèmes opérationnels sont faciles à manipuler. Ces logiciels utilisent des facilités graphiques et la plus grande partie d'entre eux comporte une certaine « intelligence », laissant aux utilisateurs très peu de choses à faire. Si d'un côté les logiciels de base sont faciles à utiliser, ils demeurent les problèmes d'utilisation des IRI, principalement celles guidées par des langages à commandes dont la syntaxe est normalement très rigide. Malgré l'existence d'un standard appelé *Common Command Language* — ISO 8770 — on trouve encore des systèmes qui adoptent une syntaxe spécifique. Il faut remarquer qu'avec le progrès technologique de l'informatique, on trouve de plus en plus de logiciels qui utilisent des facilités graphiques pour la recherche d'information, ce qui donne plus de convivialité aux IRI. Cependant, on a encore les opérateurs booléens, les opérateurs de troncature et de proximité qui exigent toujours, de l'utilisateur, une certaine maîtrise de la logique booléenne et aussi de sa syntaxe [24] . De plus, il faut que l'utilisateur connaisse le langage d'indexation de la base de données, car la requête doit être construite avec des termes qui ont été utilisés dans l'indexation de la base de données (voir la figure 1.2 dans la section 3.1). C'est la condition « sine qua non » pour qu'on puisse trouver l'information qu'on cherche dans une base de données, sauf si le SRI possède une table de synonymie ou un dictionnaire. On se rend compte que ce n'est pas suffisant de maîtriser un domaine de connaissance donné, mais il faut aussi connaître la base de donnée et son langage d'indexation.

Dans la littérature spécialisée plusieurs auteurs critiquent la rigidité des syntaxes des

[24] La syntaxe d'utilisation des opérateurs booléens est spécifique à chaque SRI, car il n'y a pas de normalisation.

langages à commande des interfaces de recherche d'information. Parmi eux on trouve [SHOW et al.] et [POLITY]. Ce genre de difficulté amène souvent les utilisateurs à suivre des cours d'accès à bases de données ou à la lecture de lourds manuels de SRI. Là on parle plutôt des IRI guidées par des langages à commandes d'une manière générale, mais il faut remarquer qu'il y a plusieurs types d'interfaces, dont on parlera par la suite.

Plusieurs systèmes d'interaction entre l'usager et l'ordinateur pour l'accès à l'information sont connus et peuvent être classés en quatre types : 1) interfaces guidées par arborescence de menus ; 2) interfaces guidées par dialogue ; 3) interfaces guidées par navigation ; 4) interfaces multimodales. Eventuellement, il est possible de trouver des interfaces que combinent les avantages des types cités ci-dessus avec d'autres de manière à améliorer la convivialité.

3.3.1 Interfaces guidées par arborescence de menus

C'est une des plus simples des interfaces, puisque l'usager ne fait que choisir et exécuter l'option qu'il faut, à chaque menu ; ces interfaces sont normalement construites sur mesure, c'est-à-dire qu'elles sont spécifiques pour chaque application ou base de données. Ainsi elles tendent à être très efficaces. C'est une sorte d'interface la plus adaptée aux usagers naïfs ou novices. Elles ont l'inconvénient d'être très rigides, car on ne peut pas construire une requête selon ce qu'on veut mais selon ce qui est offert dans les menus. Une autre contrainte est qu'on ne peut pas mettre un nombre important d'options dans un menu car on risque d'ennuyer les usagers, surtout s'ils sont expérimentés [25] .

Selon Yolla Polity ***« Les défauts des interfaces à menus ont été à maintes reprises soulignés : rigidité, lenteur du déroulement, inefficacité de la recherche multicritères (seul l'opérateur « et » est généralement disponible). Cependant, on constate que pour une utilisation occasionnelle, et pour des applications très ciblées ne mettant pas en jeu un vocabulaire très étendu, elles ont un grand succès. »*** [26] . Ces problèmes énumérés par Yolla existent, mais il faut remarquer qu'il existe des moyens de rendre plus rapide le déroulement du menu, c'est alors un problème de structuration des options placées dans les menus ; ou un déroulement meilleur, c'est alors un problème de conception de l'interface. Rudy VAN HOE et al. [27] ont fait quelques expériences sur la construction des interfaces guidées par menus et ils sont arrivés à quelques conclusions par rapport au temps nécessaire pour trouver une information :

[25] « Usagers expérimentés » c'est une dénomination donnée aux usagers que sont habitués à manipuler les SRI ou qui ont l'expertise de la recherche d'information.

[26] Yolla POLITY. " Evaluation des modes de recherche en langage naturel ". *Documentaliste - Sciences de l'Information*. 1994, vol. 31, n°. 3. p. 138.

[27] Rudy VAN HOE, R. ; Karel POUPEYE ; André VANDIERENDONCK ; et Geert DE SOETE. « Some effects of menu characteristics and user personality on performance with menu-driven interfaces ». *Behaviour & Information Technology*. 1990, v. 9, n. 1, p. 27.

1. Le temps de sélection dans un menu est plus court lorsque les systèmes sont basés sur des données structurées de manière hiérarchique ;
2. Le temps moyen par écran est une fonction croissante du nombre d'alternatives présentées dans un écran. Une explication à cette conclusion est qu'il faut un temps minimal nécessaire pour prendre une décision ;
3. Le temps nécessaire pour trouver un item d'information dans une base de données est plus court lorsque la structure est moins profonde : on trouve un item d'information plus rapidement dans une structure de menus de deux ou trois niveaux que dans une structure plus profonde.

Ainsi, si on veut construire une interface guidée par arborescence de menus où il y a 64 options différentes, il vaut mieux les mettre dans une structure de menus à deux niveaux avec 8 options chacun. Outre ces conclusions, ils conseillent aux développeurs d'interfaces à menus, de prendre le soin de mettre toujours, dans chaque menu, une option permettant de retourner au menu précédent et même au menu de début. Cette procédure donne plus de convivialité aux utilisateurs, car ils peuvent refaire le chemin lorsqu'ils se sont trompés, c'est une façon de permettre aux utilisateurs de corriger une prise de décision erronée.

Pour les applications de recherche d'information documentaire, un système appelé CANSEARCH [28] , utilise une interface guidée par arborescence de menus.

3.3.2 Interfaces guidées par dialogue

Ces interfaces ont soit un langage artificiel soit un langage naturel. On appelle les interfaces guidées « à langage artificiel », celles qui utilisent des langages construits par les développeurs d'interface ; et « à langage naturel », celles qui utilisent un langage pareil au langage parlé par les hommes pour communiquer avec un usager.

3.3.2.1 Interfaces guidées par langages artificiels

Les principales caractéristiques de ce type de langage viennent du fait qu'il est artificiellement défini, qu'il possède des règles rigides de syntaxe et qu'il utilise des expressions complexes (booléennes et/ou mathématiques).

Par conséquent, l'utilisateur novice trouve des difficultés à les utiliser. Les applications qui utilisent souvent ce type de langage sont principalement celles de la

[28] CANSEARCH est un système de recherche d'information sur la littérature concernant la thérapie du cancer dans la base Medline. Il a été développé par [POLLITT] à l'usage des novices. Il est basé sur une arborescence de menus où les usagers font la sélection des options en touchant un écran tactile. Ils n'utilisent pas de claviers. Les usagers construisent la requête en sélectionnant les mots ou les phrases présentés dans les différents menus sur l'écran. Ce type d'approche est connu comme menu en langage naturel et il utilise des bases de connaissances pour la représentation du contenu des documents. Thompson et ses collègues ont développé un ensemble d'outils pour la construction d'interfaces de recherche d'information orientée à menus en langage naturel. ALBERICO, Ralph et MICCO Mary. *Expert Systems for Reference and Information Retrieval*. Westport : Meckler, 1990. p. 96 (Supplements to computers in libraries).

recherche documentaire (langage à commande basé sur des expressions booléennes) et les interfaces guidées par SQL [29] .

En ce qui concerne les applications de la recherche documentaire ou de la recherche de texte intégral, on a fait des remarques, dans l'introduction de cette section, à cause des problèmes du manque de convivialité de ce type d'interface.

Les grands serveurs de bases de donnés (DIALOG, STN, QUESTEL, etc.) utilisent des interfaces guidées par des langages à commandes, lesquelles permettent, d'une manière générale :

- La formulation d'une recherche d'information à travers la construction d'une expression composée par l'indication de champ (titre, auteur, descripteur, etc.), de mots clés combinés par des opérateurs booléens [OR (ou), AND (et), NOT (sauf)] ;
- La formulation d'une recherche utilisant des recherches déjà faites, c'est-à-dire une formulation rappelant les recherches antérieures ;
- Une grande partie des interfaces de recherche des grands serveurs de base de données permet l'utilisation des opérateurs de voisinage, de proximité ou de distance. Ce qui permet d'établir, dans une recherche d'information sur un champ donné, l'ordre et la distance entre les mots dans un document ; Exemple : **information** ***(w)*** **retrieval** ***(w)*** **systems** => cette expression indique qu'on cherche les références que possèdent les mots *information, retrieval* et *systems* disposés à la suite (l'opérateur *(w)* indique que les mots doivent être adjacents) comme : *information retrieval systems.* On améliore la précision du résultat.
- On trouve aussi la facilité d'utiliser les opérateurs de troncature, normalement à droite, pour spécifier des préfixes, ainsi que des facilités de troncature à gauche et au milieu du mot, lesquels sont moins fréquents que celle à droite ; Exemple : biblio+ => cette expression indique qu'on cherche tous les mots avec la base " biblio ", comme résultat on peut obtenir des références avec les mots : bibliobus, bibliographie, bibliographique, bibliometrie, bibliophile, bibliophiles, bibliophilie, bibliothéconomie, bibliothèque. Ce genre de recherche peut demander un temps de réponse plus long car il est directement proportionnel au nombre de descripteurs du lexique répondant au critère de troncature. Bien qu'une consultation du lexique d'une base soit une façon sûre de lever une ambiguïté, il faut malgré tout prendre garde à l'utilisation de ce type d'opérateur. Ils provoquent beaucoup de bruit ;
- Quelques interfaces permettent d'utiliser un thésaurus comme aide à la recherche d'information ;
- Visualisation des résultats en ligne ;
- Impression des résultats en différé ;

D'une manière générale, ces interfaces utilisent comme commande un verbe, soit à l'infinitif soit à la troisième personne du singulier, suivie d'une syntaxe spécifique à chaque

[29] SQL est la sigle de Structured Query Language, langage normalement utilisé pour la recherche sur des bases de données relationnelles.

commande, selon sa fonction. Ce type d'interface est plus approprié aux usagers expérimentés et aux professionnels de l'information.

3.3.2.2 Interfaces guidées par langage naturel

La caractéristique principale de ces interfaces est d'utiliser un langage similaire à la langue parlée et d'éviter la complexité d'une syntaxe artificielle. La construction de ces interfaces est basée sur les techniques du traitement automatique du langage naturel. On peut faire une demande d'information en utilisant des requêtes formulées en langage naturel et recevoir une réponse également en langage naturel. Un avantage de ce type d'interface est qu'il n'est pas nécessaire d'apprendre un nouveau langage, on utilise son propre langage, ce qui donne à ce type d'interface plus de convivialité. De plus, il est plus facile aux usagers d'exprimer leurs besoins d'information en langage naturel que dans un langage à commande artificielle.

Selon BINOT et al., on peut distinguer trois grandes générations d'interfaces à langage naturel : a) interfaces à traduction directe ; b) interfaces à traduction avec un langage intermédiaire ; c) interfaces à modèle du discours avec des modèles de l'utilisateur.

a) Interfaces à traduction directe

D'après BINOT et al. « *Les premières interfaces développées étaient 'à traduction directe', c'est-à-dire qu'elles se composaient essentiellement d'un analyseur-traducteur transformant directement l'énoncé du langage naturel dans un formalisme adéquat pour l'application visée (par exemple un formalisme d'interrogation de bases de données)* » [30] . Pourtant, ce type d'interface présentait quelques inconvénients majeurs qui ont conduit à l'adoption d'une deuxième architecture, celle basée sur la traduction du langage naturel en un langage intermédiaire.

b) Interfaces à traduction avec un langage intermédiaire

Toujours d'après Jean-Louis BINOT et al., ***« La deuxième génération d'interfaces est basée sur la notion de " langage intermédiaire ". Ce second modèle, adopté par la grande majorité de systèmes actuels, divise le processus de compréhension en deux étapes :***

1. un analyseur général traite l'énoncé en langage naturel pour produire une représentation de son contenu sémantique dans un « formalisme intermédiaire de représentation du sens » ;
2. un interprète, dépendant de l'application visée, examine le contenu de la représentation intermédiaire et décide des opérations qu'il importe d'effectuer dans le contexte de l'application, ce qui comprend normalement la traduction de tout ou de partie de la forme intermédiaire dans un formalisme spécifique à

[30] Jean-Louis BINOT, Lieve DEBILLE, David SEDLOCK, Bart VANDECAPELLE. " Représentation Sémantique et interprétation dans une Interface en Langage Naturel ". *Le Français Moderne*. Juin, 1991, t. LIX, n°. 1. p. 56.

l'application. »Jean-Louis BINOT, Lieve DEBILLE, David SEDLOCK, Bart VANDECAPELLE. " Représentation Sémantique et interprétation dans une Interface en Langage Naturel ". Le Français Moderne. Juin, 1991, t. LIX, n°. 1. p. 57.

Un exemple de ce type d'interface est le système LOQUI [31] , une interface pour bases de données relationnelles.

c) Interface à modèle du domaine du discours avec un modèle de l'utilisateur

« Enfin on peut distinguer une troisième génération d'interfaces, caractérisée par le fait qu'elles incluent, outre un modèle explicite du domaine du discours, un modèle explicite de l'utilisateur » [32] . Ce genre d'interface offre un avantage de plus par rapport aux deux types précédents puisqu'il prend en compte le modèle de l'utilisateur, de manière à lui donner des résultats plus précis en réponse à son besoin d'information.

Un exemple de ce type d'interface est le projet Esprit MMI2, décrit par Jean-Louis BINOT et al. [33] et aussi par Samir DAMI & Geneviève LALLICH-BOIDIN [34] . Ce projet est aussi un exemple d'interface multimodale, il y a donc une brève descriptions de cette interface dans la section 3.3.4.

3.3.3 Interfaces guidées par navigation

Ces interfaces comprennent les systèmes hypertextes et hypermédias. Le mot hypertexte a été créé par Theodore H. Nelson pour désigner des écrits non-linéaires. En fait une

[31] « *LOQUI est une interface de dialogue portable et multilingue dont le premier prototype fut développé dans le cadre d'un projet de coopération internationale ESPRIT…les principaux éléments de son architecture son les suivantes : la chaîne de caractères constituant l'énoncé est d'abord soumise à un traitement lexical et morphologique permettant d'identifier un 'lexique dynamique' : liste des sens possibles de tous les mots présents ; un analyseur génère à partir de ce lexique dynamique une représentation sémantique interne ('semrep') représentant le sens de la phrase ; cette représentation sémantique intermédiaire fait l'objet d'un processus d'interprétation sémantique et pragmatique dont un des aspects réside dans l'extraction des informations appropriées de la base de données ; un processus de détermination de la réponse génère, à partir du résultat fourni par la base de données et de diverses informations relatives à l'énoncé et au contexte du dialogue, une nouvelle forme sémantique intermédiaire décrivant le contenu de la réponse à générer ; un générateur 'tactique' produit une réponse en langage naturel à partir de la forme sémantique intermédiaire qui lui est fournie ; les processus qui précèdent s'appuient sur un modèle conceptuel du domaine d'application et sur un modèle explicite de la structure du dialogue en cours et des informations liées au contexte du dialogue. On notera encore que le langage utilisé pour exprimer les formes sémantiques intermédiaires acceptées en entrées par le générateur est le même que celui des formes produites par l'analyseur, de sorte qu'il est possible de connecter directement l'un à l'autre pour faire fonctionner le système en mode paraphrase.* ». Ibidem p. 56-59.

[32] Ibidem p. 57.

[33] J.-L. BINOT, P. FALZON, R. PEREZ, B. PEROCHE, N. SHEEHY, J. ROUAULT et M. WILSON. " Architecture of a multimodal dialogue interface for knowledge-based systems ". In. : *Actes de la Conférence ESPRIT 90*. Novembre, 1990. p. 412-433.

[34] Samir DAMI et Geneviève LALLICH-BOIDIN. « An expert system for French Analysis within a Multi-Mode Dialogue to be Connected ». *RIAO 91 : Recherche d'Information Assistée par Ordinateur*. Barcelona, 1991. vol. 1. p. 431-451.

interface hypertexte permet aux usagers de lire un texte de façon non séquentielle. Ainsi un système hypertexte, à l'origine, était plutôt un système d'organisation et de structuration de texte, permettant une lecture non séquentielle du texte. La façon de stocker et de trouver l'information, dans ces systèmes, ressemble à celle utilisée couramment par l'homme, par association d'idées.

Dans un premier temps, il était très facile de trouver l'information, mais dans un petit ensemble de textes. Or, pour des grandes bases de données, ce genre de systèmes ne fonctionnait pas bien car ce n'était pas un système de recherche d'information, dans le sens pur du terme, mais plutôt un système de lecture non séquentiel. On a donc ajouté à ces systèmes une fonction de recherche d'information basée sur l'approche traditionnelle, par utilisation de mots clés composés avec des opérateurs booléens. Ainsi, le démarrage d'une recherche d'information, dans un système hypertexte, commence d'abord, soit par la formulation d'une requête de recherche d'information traditionnelle, soit à partir d'un document donné, puis on navigue dans les textes des documents. Ces nouveaux systèmes ont provoqué l'apparition des « *Hypertexts Information Retrieval* ». Ce qui a apporté quelques changements dans le champ de la recherche d'information. Deux aspects importants sont apparus avec les hypertextes : 1) une augmentation de l'interactivité entre les usagers et le système de recherche d'information ; 2) une visualisation plus riche des résultats d'une recherche. Les points remarquables de ce changement sont : 1) l'utilisateur peut avoir la possibilité de mieux connaître le contenu de la base de données ; 2) la procédure de construction d'une requête de recherche d'information sera plus dynamique à cause de l'interactivité, ce qui permet un « *feedback* » plus rapide et ainsi la réformulation de la requête.

En réalité, les systèmes hypertextes ont acquis leur faisabilité grâce à des facilités graphiques, des écrans colorés et à l'insertion de la souris comme périphérique d'un ordinateur. Une autre facilité importante, souvent utilisée par ces systèmes est la meilleure utilisation de l'écran, les systèmes opérationnels offrent la possibilité de pouvoir partager plusieurs fenêtres dans un même espace de l'écran.

Ainsi, les principales caractéristiques d'une interface de navigation hypertextuelle sont :

1. la multiplicité des parcours de lecture des informations ;
2. l'utilisation de multiples fenêtres : on peut visualiser plusieurs documents sur un même espace de l'écran. C'est à dire, on peut visualiser plusieurs documents simultanément sur un même écran ;
3. l'interactivité : les interfaces hypertextuelles offrent à l'usager plusieurs choix de parcours. Les SRI classiques, lors de la procédure de présentation des documents trouvés en réponse à une requête, n'offrent qu'un parcours séquentiel. Tandis que les interfaces de navigation hypertextuelles permettent à l'usager voir les documents dans un ordre quelconque en suivant les liens existants dans les documents trouvés, par l'ordinateur, en réponse à une requête. En effet, ces liens peuvent amener l'usager à d'autres documents hors l'ensemble de ceux qui constituent la réponse à sa requête. Cette souplesse lui permet de reformuler la requête originelle ou même

de sélectionner les documents parcourus qui lui satisfont son besoin d'information.

Il existe aujourd'hui un autre type de système nommé « systèmes hypermédias ». Selon Jean-Pierre Balpe, ils sont ***« … en effet, un ensemble d'informations appartenant à plusieurs types de médias (textes, son, image, logiciels) pouvant être lus, écoutés ou vus suivant les multiples parcours de lectures, en utilisant également la possibilité de multifenêtrage. Ce qui différencie essentiellement l'hypermédia de l'hypertexte, n'est ainsi que la nature symbolique des codages d'information utilisées. Un hypermédia n'est rien d'autre qu'un hypertexte gérant des textes supportés par des médias divers »*** [35] . C'est à dire, les systèmes hypertextes sont en fait une sorte de systèmes hypermédias, étant donné que ces derniers comprennent des informations sous forme de textes, d'images, de son. L'organisation, la structure et l'interface sont pareilles dans les deux systèmes, ce qui change c'est la nature des informations stockées et véhiculées. Exemples des systèmes hypertexte / hypermédia : GUIDE, HyperCard, Hypergate, SuperCard, SmartBook, Storyspace etc. Ce sont des systèmes dans lesquels on peut créer une application hypertexte ou hypermédia.

Parmi les exemples d'application Hypermédia, on trouve le WWW [36] qui permet la navigation dans l'Internet pour accéder aux millions d'informations existantes dans ce réseau, soit à travers la saisie d'un adresse spécifique, soit à travers l'utilisation d'un des plusieurs moteurs de recherche (« *search engines* ») existants aujourd'hui, parmi lesquels on distingue Alta Vista, Excite, Infoseek, Lycos, HotBot, etc...

Jean-Pierre LARDY définit un moteur de recherche comme étant ***« des bases de données constituées automatiquement grâce aux logiciels robots qui scrutent à intervalles réguliers les serveurs (WEB, GOPHER, FTP ou autres selon le produit) déclarés sur l'Internet. Ils indexent mot à mot les documents localisés permettant ainsi des interrogations par sujet. Selon le moteur de recherche utilisé, les recherches porteront sur : le titre ou l'entête des documents ; ou les documents complets »*** [37] . On voit là que les procédures d'indexation ressemblent à celles utilisés par les traditionnels SRI, c'est-à-dire mot à mot. Il y a, donc, les mêmes problèmes en ce qui concerne la précision des résultats de la recherche que ceux qu'on trouve dans les traditionnels SRI.

Jean-Pierre LARDY continue son explication sur les moteurs de recherche en disant que : ***« l'utilisation de ces index se veut simple et rapide : pas question d'apprendre un langage de commande pour les interroger comme pour les bases de données bibliographiques des années 80. En général la question se pose en une fois et il est impossible d'affiner petit à petit une recherche. Le volume d'information disponible fait qu'il y a presque toujours des réponses, mais au prix d'un bruit important. Pour être efficace il est utile de connaître la manière dont la question est traitée.***

[35] Jean-Pierre BALPE. *Hyperdocuments, Hypertextes, Hyermedipa*. Paris : Eyrolles, 1990, p. 18.

[36] WWW est la sigle de *World Wide Web*.

[37] Jean-Pierre LARDY. *Recherche d'Information dans Internet : outis et méthodes*.Paris : ADBS Editions. 3ème édition de mise à jour – Mai 1997, p. 60.

Malheureusement chaque moteur a son propre mode d'analyse. En général une question sera constituée d'un terme simple ou composé sans opérateurs booléens et sans caractère de troncature. L'opérateur implicite par défaut est le 'ou' (il y aura ainsi rarement des réponses nulles) et les termes sont tronqués selon des règles fonctionnant sur l'anglais. Des règles d'écriture particulière permettent d'utiliser des opérateurs d'adjacence. Les termes les plus fréquents de l'anglais sont filtrés grâce à un dictionnaire de mots vides » [38] .

Ainsi, on peut voir que ces moteurs ont apporté une certaine amélioration par rapport aux SRI traditionnels, étant donné qu'ils n'ont pas des langages de commande ou une syntaxe rigide à apprendre au préalable avant d'une consultation sur l'Internet. Cependant, cela a apporté quelques contraintes initiales parce qu'on ne pouvait pas, avec la majeure partie de ces moteurs, affiner une recherche. Toutefois, petit à petit les moteurs commencent à offrir quelques possibilités de raffinement.

Le moteur EXCITE a mis en place la possibilité de faire la recherche, en utilisant une méthode de recherche appelée *Query-By-Example* (QBE). Cette méthode consiste en faire une recherche à partir d'un document type indiqué par l'usager. Le SRI fait une recherche en utilisant les caractéristiques du document indiqué par l'usager. C'est une sorte de reformulation de requête, étant donné que l'usager fait l'indication d'un document parmi ceux trouvés par une recherche déjà faite. Dans le cas du moteur Excite, le document est une page WEB, choisi parmi les pages trouvées par une recherche faite au préalable.

Jean-Pierre LARDY complète les informations sur les moteurs de recherche : ***« La réponse à une question est une liste des adresses (URL) de sites ou de documents html en bouton hypertexte. Cette liste est en général classé par ordre de pertinence reposant sur une pondération des documents calculée à partir des critères de recherche. C'est l'application des travaux de Salton »*** [39] .

Il faut remarquer qu'aujourd'hui les choses changent très vite et les moteurs de recherche sont en pleine évolution, chaque jour apparaissent des nouveautés. De la même façon que les premiers moteurs n'avaient pas des possibilités de raffinement comme l'affirme LARDY, certains moteurs donnaient des réponses curieuses. On peut citer au moins un exemple : on a demandé, au début de l'année 1995, une recherche sur l'événement SBIA. SBIA est le sigle du Séminaire Brésilien d'Intelligence Artificielle et c'est un terme plus connu que son nom complet. La réponse à la requête a été de 180 références. C'était toutes les pages composées par des associations lesbiennes. Le moteur avait donc pris SBIA non comme un mot, mais plutôt comme une chaîne de caractères. L'explication de cette réponse, c'est que certainement, à cette époque là, il n'avait trouvé aucune page WEB sur cet événement (SBIA), et le logiciel a essayé de faire la recherche par SBI. SBI c'est une sous-chaîne de lesbienne. Quelques mois après, on a changé le moteur et ce genre de réponse n'apparaissait plus. Ainsi, il faut dire que les

[38] Jean-Pierre LARDY. *Recherche d'Information dans Internet : outis et méthodes.*Paris : ADBS Editions. 3ème édition mise à jour - Mai 1997, p. 60.

[39] Ibidem p. 61.

remarques présentées dans cette partie de la thèse — sur les moteurs de recherches — n'est valable qu'au moment de sa rédaction.

Un autre exemple intéressant d'amélioration concerne les recherches par phrases (une séquence de mots entourés par guillemets), ce qui diminue le bruit et rend les réponses plus précises. En fait, il semble que l'indexation demeure basée sur les mots, car au moment de la demande de recherche les moteurs composent la requête par l'intersection des mots de la phrase. Pour donner la réponse finale, ils font une simulation des opérateurs de voisinage ou de proximité pour ne sélectionner que les références avec les mots de la phrase fournie, dans l'ordre de cette phrase, les uns à côté des autres.

D'ici quelques mois, le panorama sera peut-être complètement différent, car les améliorations sont réalisées de manière rapide. Cette vitesse s'explique par l'agressivité du marché dans l'Internet. Il faut, donc, parler de ces outils avec précaution. Des recherches envisageant l'amélioration de ces outils sont en plein développement, notamment ceux qui sont basés sur le traitement automatique de la langue naturelle.

3.3.4 Interfaces multimodales

Avec l'arrivée des micro-ordinateurs on a vu progresser les périphériques — accessoires qui permettent entrée et sortie des informations sur un ordinateur comme les écrans, les lecteurs CD-ROM, les imprimantes, les scanners etc. — en comparaison avec le début de l'informatique où on ne connaissait que les cartons perforés, les imprimantes puis juste après, les moniteurs monochromatiques. Ce développement a permis la construction d'interfaces beaucoup plus conviviales qu'auparavant utilisant différents modes d'expression. Par ailleurs, en plus de la diversité des périphériques, la facilité d'utilisation de couleurs et de graphismes rend les interfaces beaucoup plus amusantes et conviviales. Ainsi, aujourd'hui l'utilisateur peut trouver des interfaces permettant l'interaction avec plusieurs modes d'expression, comme le pointage à travers la souris, la facilité graphique et des couleurs, les fenêtres sur l'écran, ainsi que l'interaction à travers la parole et l'écriture. A ce type d'interface on donne le nom d'interfaces multimodales. La possibilité d'interagir d'une autre façon que l'écriture évite par exemple les erreurs d'orthographe dues à la frappe erronée. Un autre avantage c'est de permettre aux usagers de choisir le mode d'expression qui s'adapte le mieux à eux. Un exemple d'interface multimodale est le projet MMI2 [40]

4 Conclusion

Le but de ce chapitre est de présenter le contexte de la recherche que nous nous sommes proposés à réaliser. Ainsi, nous ayons essayé de définir un Système de Recherche d'Information (SRI), de montrer les similitudes entre un SRI et plusieurs types de systèmes d'information, et finalement de montrer les caractéristiques et les faiblesses des SRI, notamment le manque de convivialité des interfaces de recherche d'information et la faible précision des résultats donnés par les SRI. Ces problèmes sont liés à la conception de l'interface et de l'indexation automatique (traitement de l'information) respectivement.

Nous avons vu dans ce chapitre quelques méthodes utilisées dans le développement des interfaces de recherche d'information ainsi que des procédures d'indexation. Or, il est difficile d'établir un modèle d'un système idéal de recherche d'information. On arrive plutôt à obtenir quelles sont les caractéristiques les plus souhaitables.

Améliorer la convivialité d'utilisation de ces systèmes et augmenter la précision des résultats de la recherche d'information sont les objectifs principaux. On peut ainsi envisager les caractéristiques souhaitables suivantes :

· Par rapport aux interfaces L'interface doit posséder une interaction avec l'usager de manière facile et naturelle, elle devra être construite en vue de l'usager final, ses caractéristiques les plus souhaitables sont donc :

– utilisation de mode graphique (les fenêtres, les boutons), de la souris et ainsi que du clavier pour la communication avec l'usager ;
– éviter l'utilisation de langages artificiels complexes. En cas, où un tel langage est utilisé, il faut prévoir un module plus convivial de manière à permettre aux usagers débutants d'accéder à l'information. Autrement, lorsqu'on conçoit une interface beaucoup plus conviviale plus appropriée à l'usager débutant, cette interface peut certainement gêner les usagers expérimentés, en ce cas on doit construire aussi un module qui permet à ces derniers d'accéder à l'information d'une façon plus conviviale également ;
– Avoir un mécanisme d'aide contextuel, au niveau d'utilisation de l'interface ;
– Montrer le chemin parcouru par l'utilisateur en cas d'interfaces hypertextuelles ;
– Offrir plus d'interactivité avec l'usager pour qu'il puisse ajuster ses requêtes car il est toujours difficile d'exprimer les besoins d'information dans une seule requête.

· Par rapport au traitement du contenu des documents et à l'indexation. Etant donné les problèmes posés dans ce chapitre, en ce qui concerne l'utilisation du mot simple

[40] ESPRIT II/MMI2 - Multi-Mode Interface for M-Machine Interaction C'est un système expert et il a comme but la construction d'un système de communication homme-machine capable de travailler en plusieurs modes d'expression (langage naturel, graphique, gestuel et langage à la commande). L'idée est de développer une interface interactive de manière qu'un opérateur humain puisse dialoguer avec un système expert. Ainsi, il est très important que le projet soit conçu en tenant compte d'abord de la conduite de l'utilisateur. Pour cela, font partie de l'équipe du projet des psychologues, et des informaticiens. Le système consiste en plusieurs modules " experts ". Tous les modes travaillent avec une unique base de communication représentée par une expression qui s'appelle CMR (Common formalism for representation of meaning). Une expression CMR consiste en plusieurs parties : une logique de proposition de première ordre ; des opérations d'énonciation ; des informations syntaxiques et chronologiques ; des " illocutionary force ". Sur cet aspect, le CMR est une commande de description syntaxique. Ainsi le système transforme les commandes en langage naturel en expressions CMR, et vice versa. Pour cela, le système réalise l'analyse linguistique en trois stages : 1) analyse morphologique ; 2) analyse syntaxique (son fondement linguistique et son application sous forme d'un système expert) ; 3) expressions logiques . Samir DAMI et Geneviève LALLICH-BOIDIN. « An expert system for French Analysis within a Multi-Mode Dialogue to be Connected ». In. : RIAO 91 : Recherche d'Information Assistée par Ordinateur. Barcelona, 1991. vol. 1. p. 431-451.

comme un descripteur, il faut utiliser une autre approche de façon à éviter ces problèmes. Parmi les méthodes d'indexation qu'on a vue ici, l'utilisation des syntagmes nominaux pourrait les régler. Il faut adopter une sorte de descripteur capable de porter des informations et selon Michel LE GUERN ***« la plus petite unité susceptible de servir de base à une relation référentielle autonome est le syntagme nominal »*** [41] . Les humains sont capables de distinguer en un coup d'œil, lorsqu'ils regardent un document, s'il est ou non pertinent par rapport à son besoin d'information. Certes, derrière ce processus il y a des connaissances qui ont été accumulées pendant des années. Or, lorsque les humains font le choix de documents pertinents parmi d'autres, en regardant soit les titres, soit le contenu, ils ne font que comparer les éléments référentiels existant dans ces champs avec ceux qu'ils ont défini comme leurs besoins d'information ;

On envisage ici l'importance de l'interactivité des interfaces de recherche d'information de façon à permettre aux usagers de reformuler leurs requêtes. C'est donc au moyen de cette interaction que les usagers peuvent faire un raffinement, en utilisant leur stock de savoir, et orienter les SRI vers une solution plus précise et cohérente avec leur besoin d'information.

« Le modèle est par définition ce où il n'y a rien à changer, ce qui fonctionne parfaitement ; tandis que nous voyons bien que la réalité ne fonctionne pas et s'effrite de partout ; il ne reste donc qu'à l'obliger à prendre la forme du modèle, de bon et de mauvais gré. » CALVINO,Italo. Palomar. p.108

Chapitre 2 Proposition d'un Système de Recherche d'Information

A cause des faiblesses des Systèmes de Recherche d'Information (SRI) traditionnels dues surtout au manque de convivialité des interfaces de recherche d'information et au faible précision de leurs résultats, nous proposerons un nouveau SRI.

Nous sommes arrivés à la conclusion, dans le premier chapitre, que pour améliorer les SRI, il fallait changer la manière d'un SRI d'interagir avec l'utilisateur et trouver une autre approche d'indexation qui puisse fournir des réponses plus précises aux requêtes de recherche d'information.

D'abord nous ferons une remise en cause de l'indexation automatique traditionnelle, celle qui extrait des mots clés ou simples mots pour les utiliser comme descripteurs dans un SRI traditionnel. En suite, nous exploiterons les syntagmes nominaux, et verrons comment ils peuvent être structurés ; finalement nous proposerons un nouveau SRI basé sur l'utilisation des syntagmes nominaux comme moyen d'accès à l'information.

1 SRI traditionnel : encore une remise en cause

[41] Michel LE GUERN, " Un analyseur morpho-syntaxique pour l'indexation automatique ", *Le Français Moderne*. Juin, 1991, t. LIX, n°. 1, p. 24.

Tout d'abord il faut comprendre un SRI traditionnel. La plus grande partie de ces systèmes a comme caractéristiques : a) l'utilisation de mots clés comme descripteurs ; b) l'utilisation des expressions booléennes pour la recherche d'information.

On peut dire que l'indexation des documents d'une base de données n'a du sens que pour la recherche d'information. C'est elle qui extrait des documents les informations nécessaires pour qu'on puisse les retrouver à posteriori. Ces informations sont appelées descripteurs. On a vu dans le premier chapitre qu'encore aujourd'hui la plus grande partie des SRI extrait des mots pour indexer les documents. C'est-à-dire que ces systèmes prennent les mots comme chemin pour retrouver les informations dans une base de données. Pourtant, les résultats fournis par ces systèmes, lorsqu'une recherche est faite en utilisant des requêtes composées par des mots et des opérateurs booléens (ET, OU, SAUF), ont une faible précision voire avec des taux de bruit importants.

Mais, d'un autre côté, selon la littérature consultée et aussi selon la pratique, on voit que la précision s'améliore lorsqu'on fait une recherche en utilisant des opérateurs de voisinage ou de proximité. Or, dans les deux cas on utilise les mêmes mots qui ont été indexés par la même procédure d'indexation.

La différence entre les deux cas, c'est que lorsqu'on utilise les opérateurs de proximité, le SRI fait l'appariement entre les mots extraits des documents et ceux qui se trouvent dans la requête, en utilisant aussi des informations sur la localisation des mots dans chaque document en tenant compte de leur ordre de précédence, les uns par rapport aux autres.

Selon Georges Van Slype, ***« l'indexation se définit comme l'activité consistant à représenter le contenu d'un document ou d'une question de manière analytique, c'est-à-dire à en recenser les concepts et/ou les mots. »*** [42] . Selon la dernière partie de cette définition, il est question de recenser soit les concepts et les mots, soit les concepts ou les mots. On peut accorder que les concepts, extraits d'un document, puissent représenter son contenu. Mais une liste de mots ne peut représenter que l'ensemble des unités lexicales existant dans le document.

Selon Michel LE GUERN, ***« Pour que le descripteur remplisse sa fonction, qui est de mettre en relation un objet du monde — une entité extralinguistique — avec le document qui apportera des informations sur cet objet, il faut que le descripteur soit un signe indiciaire »*** [43] . LE GUERN prend la terminologie de Peirce pour distinguer le mot de la langue (légisigne symbolique rhématique dans la typologie peircienne) de l'occurrence de ce même mot dans le discours (caractérisé comme un sinsigne indiciaire rhématique).

Or, un mot en tant qu'unité du lexique ne signifie que des propriétés. Il ne fait aucune référence à l'univers du discours. Tandis que dans le texte, le mot fait partie d'un contexte

[42] Georges VAN SLYPE. *Les langages d'indexation : conception, construction et utilisation dans les systèmes documentaires.* Paris : Les éditions d'Organisation, 1987. p. 21.

[43] Michel LE GUERN. « Un analyseur morpho-syntaxique pour l'indexation automatique ». *Le Français Moderne*. Juin, 1991, t. LIX, n°. 1, p. 23.

qui lui donne un sens spécifique. À ce moment là, il fait partie des signes indiciaires.

Lorsque l'auteur rédige un texte il rassemble les mots dans un ordre, de manière à décrire un fait, un objet du monde réel. C'est-à-dire que les mots passent d'un état où ils ne désignent que des propriétés à un autre état où ils désignent des sens spécifiques, en faisant partie du discours, en composant des unités d'information.

Il est vrai qu'il y a des SRI qui gardent aussi, au moment de l'indexation automatique, des informations sur la localisation du mot dans le texte, ce qui permet à l'interface d'offrir la facilité d'une recherche en utilisant des opérateurs de voisinage ou de proximité. Ce genre de facilité donne plus de précision aux résultats d'une recherche d'information. On peut donc imaginer que lorsqu'on garde les coordonnées d'où les mots ont été extraits dans les textes, on garde la liaison, entre les mots, construite par l'auteur.

La procédure d'indexation, utilisant l'extraction des mots, fait une démarche inversée par rapport à celle de l'auteur lorsqu'il écrit le document. La liste de mots — résultat de l'indexation automatique — est un ensemble de mots isolés les uns par rapport aux autres, ils ne désignent donc que des propriétés. Cette liste ressemble à un sous-ensemble d'un dictionnaire. Pourtant, elle peut être représentative si on l'utilise en conjonction avec les coordonnées (ou les positions) de chaque mot dans le document. En gardant les coordonnées de cette manière, on garde aussi leur contexte.

Le but de l'indexation est d'extraire des informations d'un document pour qu'on puisse le représenter de manière plus précise et de permettre aux utilisateurs de le retrouver à posteriori. Puisqu'on cherche des informations et pas des mots, l'indexation alors devrait extraire des unités d'information. On revient à la conception du descripteur de M. LE GUERN.

Quand on fait une recherche d'information en utilisant des expressions booléennes, on ne tient compte d'aucune liaison entre ces mots, les uns par rapport aux autres, ni même d'une indication d'ordre de précédence entre eux. En réalité, l'expression booléenne, est une procédure de sélection de documents basée simplement sur l'existence ou non de mots qui sont dans l'expression, à l'intérieur de chaque document. Par contre, lorsqu'on utilise des opérateurs de proximité, on spécifie de manière indirecte l'ordre de précédence et le contexte du mot, au moins leur voisinage. Ainsi, un SRI traditionnel ne peut pas être appelé comme tel, mais il serait plus précis de l'appelé Système de Recherche de Mots. Pour illustrer le problème d'utilisation des opérateurs booléens, on présentera l'exemple suivant :

Si on veut trouver des documents sur *le développement de recherches scientifiques et techniques par contrat avec des entreprises privées*, utilisant un SRI guidé par langage de commande, on doit construire une expression du type : développement ET recherche ET scientifique ET technique ET contrat ET entreprises ET privés. Cette requête ne garantissant pas que le SRI ira trouver des documents concernant le sujet : *le développement de recherches scientifiques et techniques par contrat avec des entreprises privées.* La seule garantie c'est qu'il va retrouver tous les documents qui ont les mots *développement, recherches, scientifiques, techniques, contrat, entreprises et privées* mais sans aucun rapport avec la séquence des mots ni avec la proximité de chacun d'eux, les uns par rapport aux autres. Ainsi, le SRI peut ramasser des documents

qui ont le mot *contrat* dans un paragraphe parlant de campagne publicitaire, le mot *développement* dans un autre et la phrase *recherches scientifiques et techniques* dans un autre paragraphe et ainsi de suite. Dans ce contexte, il peut ramasser de documents qui parlent de contrat publicitaire dans une entreprise intéressée par le développement d'une campagne publicitaire sur les recherches scientifiques et techniques dans les entreprises privées. Et on peut également trouver ce qu'on veut, soit des documents qui parlent aussi des *développements de recherche scientifique et technologique par contrat avec des entreprises privées*.

On arrive donc à deux conclusions : 1) une recherche par expressions booléennes ne peut pas assurer une réponse avec une bonne précision, il y aura toujours un taux de bruit plus ou moins important, dépendant du volume d'information de la base de données ; 2) un utilisateur aura du mal à formuler précisément son besoin d'information en utilisant seulement une combinaison de descripteurs et des opérateurs booléens dans une requête. Ou de façon inverse, un SRI aura du mal à comprendre exactement ce que l'utilisateur veut, en analysant simplement des expressions booléennes.

2 Proposition d'un nouveau SRI

Etant donné les problèmes concernant l'utilisation des SRI, discutés dans le premier chapitre et aussi dans le début de ce chapitre, il faut trouver une solution capable de résoudre ces problèmes. Pour cela, il faut proposer un nouveau modèle d'indexation automatique ainsi qu'une nouvelle interface de recherche d'information.

2.1 Proposition d'une approche pour l'indexation automatique

Le but de cette proposition est de construire un SRI capable d'offrir aux usagers une façon plus agréable et conviviale de faire une recherche d'information, et aussi de donner des réponses plus précises à leurs demandes d'information.

En ce qui concerne la Procédure de Traitement d'Information (PTI), on a montré les problèmes entraînés par l'utilisation de mots comme descripteurs et les causes de ces problèmes. Ainsi suivant l'indication faite dans la conclusion du premier chapitre et tenant compte des raisons qui amenaient le groupe SYDO à proposer les syntagmes nominaux, ainsi que les concepts qui y sont liés, nous adopterons cette approche pour la proposition du nouveau SRI.

Mais, dans quelle mesure l'utilisation des syntagmes nominaux comme descripteurs peut-elle apporter la solution désirée ou supprimer les problèmes présentés par les SRI traditionnels ?

Tout d'abord, selon ce qu'on a montré dans le chapitre précédant et au début de celui-ci et conformément à ce que Michel LE GUERN et Richard BOUCHÉ ont affirmé dans leurs travaux, au sein du groupe SYDO, on est convaincu que le descripteur doit être un signe doté de valeur référentielle. Les syntagmes nominaux ne sont pas des signes sans référence. Bien au contraire 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 sont plus seulement des ensembles de prédicats ou de propriétés.

Dans le syntagme nominal, chaque mot a son rôle bien défini, avec une signification plus spécifique. M. LE GUERN ainsi que Richard BOUCHÉ ont défini le syntagme nominal comme étant la plus petite unité du discours porteuse d'une valeur référentielle. On peut donc dire, que le syntagme nominal est la plus petite unité d'information dans un texte.

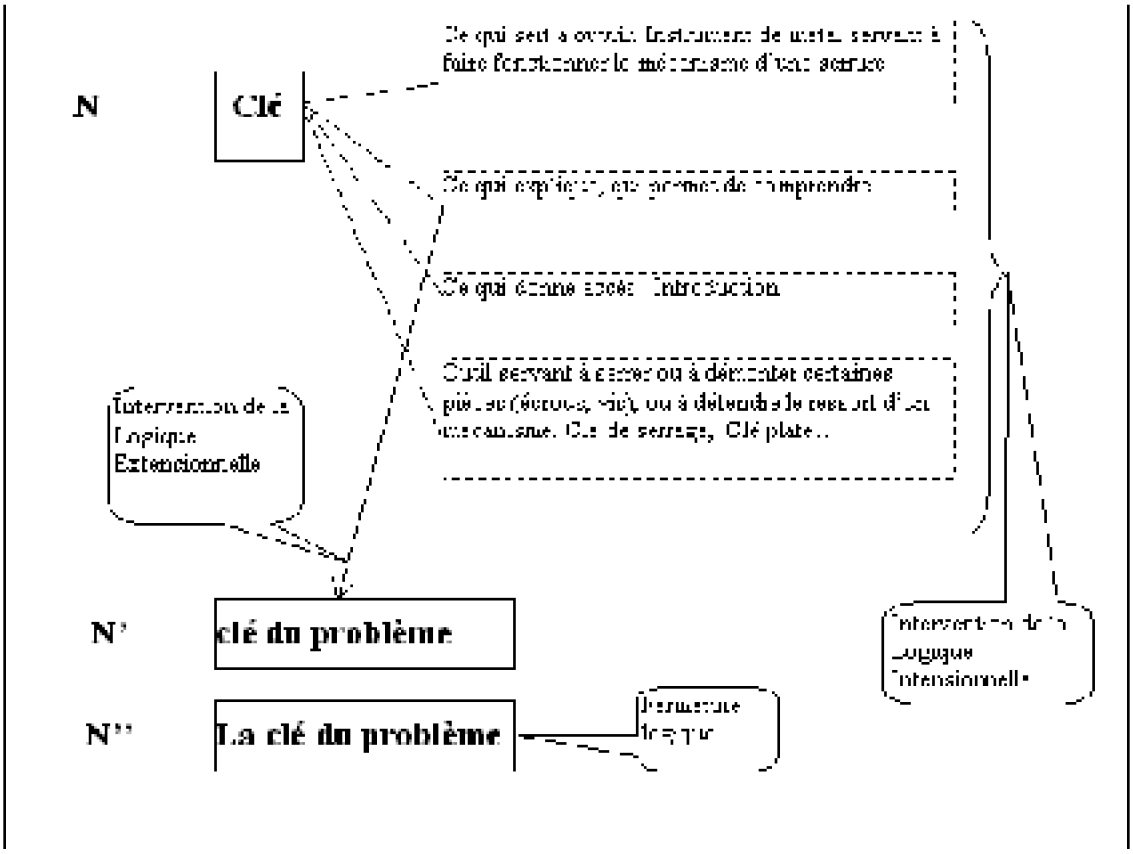
Cette opposition entre les mots en tant qu'unité du lexique, hors discours, et les mots qui font partie du discours, de manière générale et du syntagme nominal de façon plus spécifique, est bien caractérisée par Michel LE GUERN. D'abord le mot, en tant que mot de la langue, qu'unité du lexique, est au 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 c'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.

Michel LE GUERN dit que le syntagme nominal est la mise en œuvre de deux organisations logiques différentes. Au niveau N, on a l'intervention de la logique intensionnelle. La logique intensionnelle est une ***« logique sans référentiel et sans classe, constituée de relations et de propriétés envisagées indépendamment de quelque objet que ce soit »*** [44] . Tandis qu'au niveau N', il relève de la logique extensionnelle, car il 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.»*** [45] . On appelle ce qui appartient au niveau N des prédicats libres et ce qui appartient au niveau N' des prédicats liés. Le niveau N'' correspond à la fermeture logique du syntagme nominal étant donné l'ajout d'élément déterminant, c'est la complète référence à un objet donné.

Dans la figure 2.1, nous essayons de montrer la distinction entre l'intervention de la logique intensionnelle et celle de la logique extensionnelle dans la construction d'un syntagme nominal. Dans cette figure on voit de manière résumée la construction d'un syntagme nominal. C'est la manière dont l'auteur décrit un objet ou un fait. Lorsqu'il prend les mots et les rassemble dans une phrase, ces mots prennent des sens spécifiques. Les mots en tant qu'unités du lexique sont des ensembles de propriétés, ils ont une panoplie de significations possibles, ce qui donne du choix. Ils relèvent d'une logique intensionnelle. Lorsqu'ils prennent une signification précise dans l'univers du discours, ils relèvent d'une logique extensionnelle.

[44] Michel LE GUERN. « Un analyseur morpho-syntaxique pour l'indexation automatique ». *Le Français Moderne*. Juin, 1991, t. LIX, n°. 1, p. 28.

[45] Ibidem p. 28.



La logique intensionnelle est caractérisée par des éléments libres, généraux, qui peuvent désigner une ou plusieurs choses selon leurs propriétés. Tandis que la logique extensionnelle fait juste le contraire, elle caractérise les éléments qui ont des sens spécifiques, ils désignent une classe d'objets.

Il est clair maintenant que la procédure d'indexation automatique traditionnelle, lorsqu'elle extrait les mots pour construire le fichier INDICES, fait la démarche inverse de celle de l'auteur. C'est-à-dire qu'on part d'un état où on avait l'intervention de la logique extensionnelle vers un autre état où on revient à avoir l'intervention de la logique intensionnelle. En d'autres mots, la liste perd le contexte ou le sens spécifique existant dans le document donné par son auteur au moment de son écriture. Ainsi, l'indexation traditionnelle c'est le basculement de la logique extensionnelle vers la logique intensionnelle. Or le rôle de l'indexation n'est pas cela : elle devrait être une procédure de repérage de pistes capables de caractériser les documents pour la recherche d'information. Pourtant, la procédure d'indexation par mots n'extrait pas des pistes pour retrouver les documents puisque les mots extraits perdent leur valeur référentielle au

moment de leur détachement du texte. Les pistes, résultat de la procédure de l'indexation en tant que valeur référentielle devraient être fidèles à leur origine pour qu'on puisse retrouver les documents à posteriori.

Ainsi, les syntagmes nominaux peuvent jouer le rôle de descripteurs dans un SRI car ils constituent des unités d'information, ils font référence à des objets de la réalité extralinguistique de l'auteur, ils gardent les sens originaux que l'auteur leur a donnés lors de l'écriture du document. De plus, les syntagmes nominaux enlèvent les ambiguïtés qu'on a lors de l'utilisation de mots comme descripteurs. Or, on a encore l'ambiguïté due au domaine de la connaissance puisqu'on peut trouver le même syntagme nominal dans des domaines différents. Mais ce genre de problème peut être résolu au fur et à mesure qu'on fait la séparation des documents, en créant des bases de données spécialisées par domaines bien précis.

2.2 Proposition d'une interface de SRI

En adoptant l'approche conçue par Michel LE GUERN, reste la question : comment utiliser les syntagmes nominaux dans une interface de recherche d'information ? Est-ce qu'on peut utiliser les mêmes interfaces qu'on utilisait dans les SRI traditionnelles, celles guidées par un langage à commande ?

Du point de vue technique il semble que rien n'empêche de les utiliser comme on utilisait les mots dans des expressions booléennes. Cependant, nous pensons que l'usage de ces expressions doit changer maintenant avec les syntagmes nominaux. Etant donné que ceux-ci désignent des substances, ils sont dotés de valeurs référentielles, la construction des expressions booléennes utilisera beaucoup plus les opérateurs OU et SAUF que l'opérateur ET. En fait, cette remarque est une simple supposition sans aucune corroboration pratique. Pour corroborer cette supposition il faut faire des observations pratiques avec les usagers en utilisant un SRI avec cette approche.

Une autre remarque qu'on peut faire à cet égard, c'est par rapport à la saisie d'un syntagme nominal ou à la saisie d'une combinaison des syntagmes nominaux et des opérateurs booléens. Ce qui peut affecter la convivialité de l'interface lorsqu'on a des longues requêtes (des syntagmes nominaux longs ou des combinaisons de syntagmes nominaux longs avec des opérateurs booléens). L'ennui peut arriver non seulement de la longueur de la requête, mais aussi d'éventuelles fautes d'orthographe, fréquentes dans une saisie de textes longs. Ainsi, il nous semble que les syntagmes nominaux peuvent, dans ce contexte, améliorer la précision des résultats fournis par ces interfaces, mais ils ne peuvent certainement pas améliorer la convivialité de ces interfaces.

Geneviève LALLICH-BOIDIN dit en sa thèse : ***« L'interrogation d'un fonds documentaire où chaque document est représenté par une liste de syntagmes nominaux se doit de partir d'une question formulée en langue naturelle. De cette question, en sont extraits les syntagmes nominaux qui seront comparés à ceux présents dans la base. Sans entrer dans les détails d'une stratégie d'interrogation, stratégie qui est encore à l'état d'étude, l'on peut cependant avancer que, la représentation choisie est très pertinente pour une recherche documentaire, malgré quelques handicaps. »***

Il est évident que les besoins d'information d'un usager sont ou doivent être exprimés en langue naturelle. Dans la procédure traditionnelle, à partir de cela on élabore une requête laquelle est soumise à un SRI pour trouver l'information qui on a besoin. Cette requête n'est toujours pas écrite en langue naturelle. Au contraire, elle est normalement exprimée dans un langage artificiel guidé par commandes. En ce qui concerne l'usage des syntagmes nominaux dans une procédure de recherche d'information, il n'y a aucune exigence que la requête soit faite en langue naturelle. A mon avis nous pouvons exploiter les syntagmes nominaux comme pistes pour la recherche d'information autant dans une interface guidée par un langage naturel ou un langage artificiel que dans celle guidée par menu.

Pourtant, si LALLICH-BOIDIN considère que les mots « question » et « requête » sont la même chose, cela suscite une réflexion plus profonde. En ce cas, elle me semble proposer tout simplement une procédure de recherche d'information où l'usager exprime son besoin d'information dans une requête en langue naturelle. En suite, dans sa conception, l'ordinateur extrait une liste de syntagmes nominaux présents dans la requête et la compare contre une liste pareille extrait du fonds documentaire. Or, nous savons qu'une chose, un objet peut être nommé par un ou plusieurs syntagmes nominaux. Ainsi, dans cette proposition on peut donc avoir du silence dans les résultats de la recherche. C'est-à-dire, bien que des informations soient présentes dans une base de données elles ne seront pas trouvées à cause d'utilisation de SN ou descripteurs équivalents ou synonymes à ceux présents dans les documents de la base de données. Par exemple, si l'usager cherche les documents qui parlent de « la recherche de documents », disons que dans le fonds documentaire les documents traitent plutôt de « la recherche d'information » que de « la recherche de documents ». Bien que les deux syntagmes nominaux désignent la même chose, le terme le plus utilisé dans la littérature spécialisée est « la recherche d'information ». Dans cet exemple, le SRI proposé par LALLICH-BOIDIN, ne donnera sûrement pas la bonne réponse. Il y aura du silence.

Une autre réflexion que la citation présentée suscite est : un usager, arrive-t-il à exprimer son besoin d'information du premier coup devant une base de données quelconque ? On sait, par la pratique, qu'un usager, un spécialiste ou un technicien en information n'arrive à exprimer leur besoin d'information, avec précision, qu'après l'absorption de quelques connaissances sur la base de données. C'est seulement après quelques essais de recherche d'information sur une base de données qu'on peut s'exprimer avec sûreté, en élaborant une requête, ce qu'on veut. Normalement l'usager a une idée floue de ce qu'il veut trouver. Cela peut arriver par deux raisons : 1) il ne sait normalement pas quels sont exactement les descripteurs qui expriment le mieux son besoin d'information ; et 2) il ne connaît pas comment la base de données a été indexée. Il y a évidemment plusieurs types d'usagers. Il y a ceux qui ne connaissent pas une base de données et même un SRI, il y a d'autres qui connaissent déjà un SRI mais ils ne connaissent pas la base de données qu'ils vont consulter. Et, il y a d'autres qui connaissent aussi bien le SRI que la base de données qu'ils vont consulter. Ces derniers sont des usagers expérimentés ou ce sont des spécialistes. Ces usagers savent ce qu'ils veulent et savent aussi, dans la plus grande partie des situations comment s'exprimer devant un SRI. Cependant, ils ne sont pas la majorité des usagers, bien au contraire ils

sont la minorité de l'ensemble d'usagers. On ne peut pas proposer un SRI envisageant seulement cette catégorie d'usagers.

De ce qui nous avons appris dans la littérature spécialisée et aussi dans la pratique, le processus de recherche d'information devrait être une activité interactive entre l'usager et le SRI. Mais, pourquoi cela ? C'est parce qu'il faut une participation majeure de l'homme dans le processus de recherche d'information. Pour que l'usager puisse faire une demande de recherche d'information de manière correcte, il faut qu'il connaisse bien la base de données, il faut qu'il connaisse bien l'indexation de la base de données. Ces connaissances permettent aux usagers de bien élaborer leurs requêtes de manière à obtenir de bonnes réponses en satisfaisant leur besoin d'information. Il faut donc que le SRI soit interactif de façon à faciliter aux usagers l'apprentissage du langage d'indexation de la base de données.

Nous sommes d'accord que nous pouvons utiliser une liste de syntagmes nominaux extraits d'un ensemble de documents comme moyen d'accès à ces documents. Cependant nous ne pouvons pas affirmer que cela soit *« une représentation très pertinente pour une recherche documentaire* ». Il est vrai qu'une liste de syntagmes nominaux peut apporter plus de précision dans une recherche documentaire qu'une liste de mots isolés. Mais, pour cela il faut que le fonds documentaire soit centré sur un domaine de la connaissance le plus spécifique possible. Si le fond documentaire porte sur un large domaine de la connaissance ou s'il est multidisciplinaire, on trouvera encore des syntagmes nominaux ambigus, surtout à ceux de premiers niveaux.

Il faut encore prendre en compte les caractéristiques d'un fonds documentaire. L'utilisation des syntagmes nominaux comme moyen d'accès aux documents est plus appropriée lorsqu'on a des textes entiers des documents dans la base de données. Dans un autre côté, on n'est pas sur si l'utilisation des syntagmes nominaux est pertinente comme moyen d'accès à une base de données bibliographique. Une base de données bibliographique est composé d'informations descriptives des ouvrages, des articles (p.ex. : titre, auteur, éditeur, résumé etc.). Ainsi les seuls champs qui peuvent aider l'indexation, en apportant des pistes pour la recherche d'information, sont le titre, le sujet et le résumé. Les textes des documents d'une base de données bibliographiques ne sont donc pas dans la base. Pour qu'une liste de syntagmes nominaux soit pertinente il faut que le champ résumé soit une synthèse fidèle du document original. C'est-à-dire, il faut que les principaux syntagmes nominaux existants dans le document et qui peuvent jouer le rôle de descripteur, soient présents aussi dans le résumé. Le titre lui-même est un syntagme nominal par définition. Mais le titre ne peut pas contenir tous les syntagmes nominaux possibles de représenter des pistes qui puissent faciliter la recherche d'information. Ainsi, le modèle d'indexation automatique basé sur les syntagmes nominaux est bien adapté aux bases de données texte entier (« full text ») car elle peut fournir tous les syntagmes nominaux qui ont été inclus dans le document par son auteur. Ce qui n'arrive toujours pas dans le résumé d'une notice bibliographique.

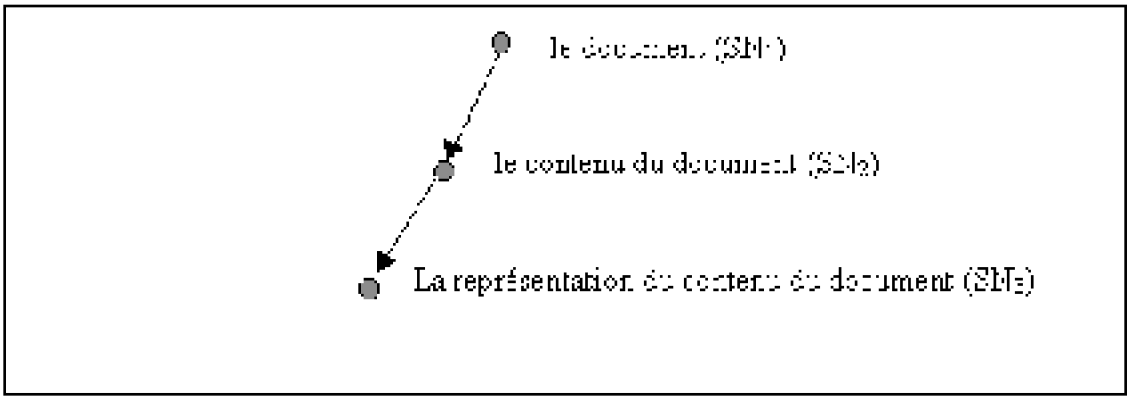
De plus, pour la réussite d'usage des syntagmes nominaux comme moyen d'accès aux documents dans un SRI, il faut qu'ils soient organisés dans une structure qui facilite la recherche d'information. C'est-à-dire, il n'est pas suffit de remplacer les descripteurs représentés par des mots, par ceux représentés par des syntagmes nominaux. Il faut

proposer non seulement un modèle d'extraction de syntagmes nominaux, mais il faut aussi ajouter une proposition d'organisation des syntagmes nominaux et un modèle d'interface de recherche d'information. C'est-à-dire, il faut proposer un nouveau modèle de SRI.

Par ailleurs, les syntagmes nominaux ont une organisation naturelle dans la mesure où ils ont un rapport d'emboîtement les uns avec les autres, ce qui permet de les rassembler dans une structure en arbre. Cette caractéristique permet de construire une interface navigationnelle capable d'exploiter les données au moyen de la navigation dans sa structure arborescente.

Pour montrer cette caractéristique on présentera un exemple, dans la figure 2.2, d'un syntagme nominal de troisième niveau [46] .

Exemple : « La représentation du contenu du document »



Dans la Figure 2.2, nous avons trois syntagmes nominaux, enchaînés en trois niveaux différents. Comme SN_1 [47] , nous avons *le document* qui a été extrait d'un SN_2 [48] *le contenu du document* lequel à son tour a été extrait d'un SN_3 *La représentation du contenu du document*. Le rassemblement de tous les syntagmes nominaux d'une base de données permettra de construire une structure arborescente. Cette structure est bien appropriée à la construction d'une interface navigationnelle. Dans la mesure où ce genre d'interface n'exige pas la maîtrise d'un langage de commande, ni l'utilisation d'opérateurs booléens pour demander la recherche d'information ni la saisie d'une expression trop grande, il nous semble que cette interface tend à être plus conviviale.

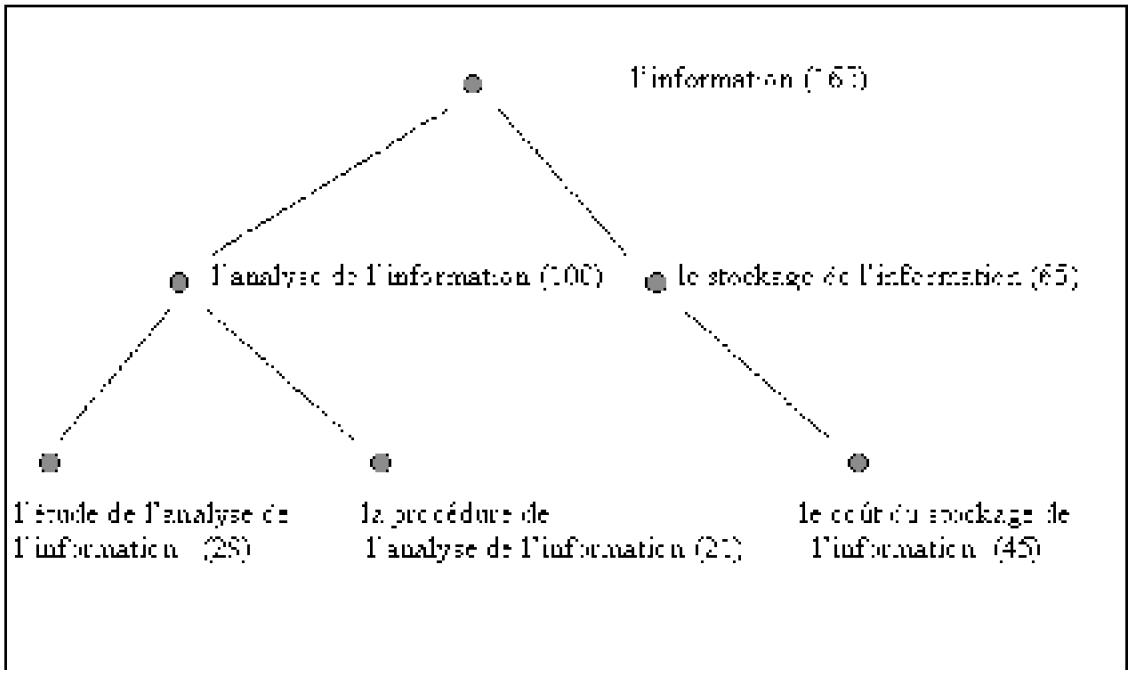
Pour illustrer comment on peut construire une telle interface, nous allons montrer dans la figure 2.3, un sous-ensemble d'une structure de syntagmes nominaux. Le nombre qui apparaît entre parenthèses est le nombre de documents d'où chaque syntagme nominal a été extrait. D'abord, le syntagme nominal de premier niveau *l'information* a été

[46] On utilise le mot niveau pour indiquer l'ordre d' extraction des syntagmes nominaux. Le syntagme de plus haut niveau c'est le syntagme le plus grand et le syntagme de plus bas niveau c'est le dernière syntagme nominal extrait, soit le syntagme nominal le plus simple. En effet la grandeur du niveau est inversé par rapport à l'ordre d'extraction.

[47] Syntagme Nominal de premier niveau.

[48] Syntagme Nominal de deuxième niveau.

extrait de 165 documents. C'est-à-dire ce syntagme nominal est présent dans 165 documents. Ensuite, on a 100 documents d'où on a extrait le syntagme *l'analyse de l'information*. Ce syntagme fait partie d'un syntagme nominal de troisième niveau, *l'étude de l'analyse de l'information,* extrait de 29 documents. Il fait partie aussi d'un autre syntagme nominal de troisième niveau, *la procédure de l'analyse de l'information,* extrait de 21autres documents restants. C'est-à-dire qu'il y a 50 (21 + 29) documents dont le syntagme nominal *l'analyse de l'information* fait partie d'un syntagme de troisième niveau. Il existe donc 50 documents qui le contiennent tout seul mais sans être lié à un autre syntagme de plus haut niveau. Dans cet exemple on peut voir clairement le processus de raffinement d'une recherche d'information, lorsqu'on descend la structure de syntagmes nominaux.



L'idée générale, c'est que l'usager fournisse à l'interface un mot qui représente son besoin d'information ou plutôt qui fait partie de son besoin d'information. A partir de ce mot, l'interface cherche et montre à l'écran, tous les SN_1 ayant le mot fourni par l'utilisateur comme leur centre de syntagme. Certainement qu'à ce moment là, on aura beaucoup de bruit, mais c'est à l'utilisateur de choisir le bon syntagme nominal pour suivre la procédure de recherche d'information. Une fois choisi le bon syntagme nominal, il peut demander à l'interface : a) soit de montrer les documents d'où ce syntagme a été extrait ; ou b) un raffinement à travers la recherche des SN_2 d'où ce SN_1 a été extrait. L'interface répète la même procédure, elle cherche les SN_2 et les montre immédiatement à l'écran et c'est à l'usager de décider l'arrêt, la suite de la navigation sur l'arbre des syntagmes ou de retourner à un niveau précédant ou au premier niveau, pour refaire la stratégie de recherche d'information. Le dernier niveau de syntagmes nominaux peut être défini soit par la programmation du système d'extraction des syntagmes nominaux, soit par une définition au préalable de la procédure d'indexation. Cependant, il semble qu'il y ait une limite naturelle du niveau des syntagmes nominaux, lequel est déterminé par la capacité humaine d'exprimer. En gros, la limite est vers le niveau 5.

L'approche proposée ici offre aux usagers une aide à la formation de la requête sans utiliser un langage de commande ou des opérateurs booléens. L'interface proposée est interactive, ce qui permet aux utilisateurs de faire des corrections de route, en améliorant leurs requêtes. C'est-à-dire que ce ne sont pas l'ordinateur ou le SRI qui font l'interprétation de la requête des usagers, ce sont eux qui conduisent la recherche d'information, ce qui donne beaucoup plus de précision aux réponses à une requête de recherche d'information.

Ainsi nous venons de voir ici, que l'interactivité peut aider à améliorer l'exactitude, dans la mesure où cela permet aux utilisateurs l'opportunité d'évaluer une réponse et de reformuler leur requête de recherche d'information.

3 Conclusion

La proposition présentée utilise une approche totalement différente de celle du SRI traditionnel. Il ne s'agit pas seulement d'une nouvelle interface ou d'un nouveau SRI, mais aussi d'une technique nouvelle pour le traitement de l'information qui intègre aussi bien un modèle d'indexation que celui d'une interface de recherche d'information. C'est-à-dire, c'est une approche unique rassemblant l'indexation et l'interface de recherche d'information. Cela constitue donc une proposition d'un SRI complet. L'approche proposée possède quelques points qui peuvent lui donner des avantages importants en comparaison avec les SRI traditionnels. Ces points sont les suivants :

1. L'approche permet à l'usager la possibilité de se promener dans la structure arborescente des syntagmes nominaux. Ce fait aide l'usager à trouver le syntagme nominal qui exprime le mieux son besoin d'information. C'est une sorte d'aide à la construction d'une requête. Etant donné que la procédure est interactive, elle permet aussi la reformulation d'une requête, ou mieux, de la stratégie de recherche d'information ;
2. L'approche proposée fait disparaître les problèmes d'ambiguïté inhérents aux mots puisque la procédure d'indexation proposée maintient les unités d'information telles qu'elles sont dans les textes. C'est-à-dire que l'ensemble de pistes ou de descripteurs extraits par ce modèle est plus fidèle au texte original que celle de l'indexation traditionnelle ;
3. Cette approche n'utilise pas et ne dépend pas d'une interface guidée par un langage de commande. Ainsi, elle n'utilise pas une requête, du point de vue formel, pour demander une recherche d'information. C'est-à-dire qu'on n'utilise pas des expressions booléennes. La conséquence de cette approche c'est que l'usager n'aura pas besoin de lire ou d'apprendre un langage de recherche d'information ni de savoir utiliser les opérateurs booléens ;
4. L'approche est basée sur l'interaction avec l'usager. C'est lui qui maîtrise et oriente la recherche d'information et non l'ordinateur.

Ces points présentés plus haut indiquent qu'on peut construire un SRI avec une interface qui n'exige pas trop de connaissances, au préalable, de l'usager. Puisque l'approche

proposée ne dépend pas d'un langage de commande et qu'elle n'utilise pas d'opérateurs booléens, on peut construire une interface plus conviviale. L'utilisation des fenêtres dans l'écran, des boutons et de couleur combinée avec un pointeur comme la souris, peut rendre l'interface plus conviviale.

Selon Pierre LÉVY, on est dans un nouvel espace, c'est l'espace du savoir, où on doit valoriser et soutenir les qualités et compétences humaines. On ne peut pas tout automatiser, les humains sont les seuls non automatisables. Dans ce contexte, notre proposition est bien encadrée, car nous nous rendons compte des difficultés des SRI traditionnels et nous changeons complètement l'approche traditionnelle de traitement de l'information. Dans l'approche traditionnelle, l'utilisateur ne fait que la demande d'information à travers des requêtes, et les seuls à travailler ensuite sont le SRI et l'ordinateur. Nous proposons de mettre les humains dans la procédure de recherche d'information où ils participent plus activement de manière interactive avec le SRI et l'ordinateur. C'est à l'usager de trouver l'information qui satisfasse son besoin d'information. C'est lui qui oriente et maîtrise la procédure de recherche d'information. L'ordinateur et le SRI ne font que suivre les orientations de l'usager. Dans cette approche il n'y a donc pas l'effort que les traditionnels SRI avaient pour interpréter et chercher l'information demandée par les usagers au moyen des requêtes avec des expressions booléennes.

L'approche adoptée peut créer une nouvelle sorte de Systèmes de Recherche d'Information, celui des Systèmes de Recherche d'Information Assistée par Ordinateur – SRIAO, puisqu'elle est caractérisée par l'interaction du SRI avec l'usager et aussi parce que ce n'est pas l'ordinateur qui fait la recherche d'information tout seul, il la fait avec l'aide et sous la maîtrise de l'usager.

Deuxième Partie : la maquette d'un SRIAO

« Vous combattez et détruisez toutes les erreurs ; mais que mettez-vous à leur place ? » Deffand (Marie, marquise du), Lettre à Voltaire.

Chapitre 3 Construction d'une base de données texte plein indexées par SN

1 Constitution du corpus pour la construction de la base de données

Avant de commencer la construction de la base de données et de la maquette proprement dite, des critères pour le choix du corpus ont été établis en tenant compte du temps disponible pour achever le travail, des conditions nécessaires pour le maîtriser et le réussir, et en considérant la durée du cours de DEA, c'est-à-dire environ six (6) mois à temps partiel et deux (2) mois à temps plein. Pour la construction de la base de données proposée il a fallu donc, mettre en œuvre les étapes suivantes : (a) établissement de critères pour le choix du corpus ; (b) traitement préalable des articles ; (c) extraction des syntagmes nominaux.

1.1 Etablissement de critères pour le choix du corpus

Les critères relatifs aux conditions nécessaires pour maîtriser ce travail sont donc les suivants :

1. Taille du corpus Pour une expérimentation de recherche d'information nous avons trouvé quinze articles (voir Annexe A), constituant le nombre minimal susceptible d'être traité dans la limite du temps disponible.
2. Taille des articles Les articles choisis ont de trois à cinq pages, dans les fichiers format Word 6.0a, police Times New Roman, style normal, taille 11. Cependant nous avons rencontré des difficultés à les trouver dans cette taille, ce qui a amené à supprimer quelques paragraphes tout en ayant le soin de ne pas perdre les syntagmes nominaux importants.
3. Le domaine de connaissance du corpus Un système de recherche d'information utilisant des syntagmes nominaux sera plus performant s'il travaille sur un domaine bien défini. Cette restriction est nécessaire car il faut travailler sur un corpus homogène par rapport à l'ensemble des syntagmes nominaux, avec un minimum d'ambiguïtés, de façon à ce qu'on puisse les organiser sous forme d'arborescence. Selon MINSKY : « Dans le langage naturel, les ambiguïtés ne découlent pas seulement du fait que les mots peuvent être regroupés de diverses façons, mais encore de ce que chaque mot peut avoir plusieurs sens différents... »Marvin MINSKY. Semantic Information Processing. Cambridge, Mass. : M.I.T. Press, 1969, p. 18, cité par Hubert L. DREYFUS. Intelligence Artificielle : mythes et limites. Flamarion, 1984, p. 95. Ainsi, la définition du domaine du corpus d'une base de données est d'une importance capitale pour que les résultats de recherche soient plus précis. Pour les bases de données multidisciplinaires la bonne solution serait plutôt de les partager en plusieurs bases de données regroupées par domaines de connaissances. Nous avons donc ainsi choisi pour ce travail le domaine des Sciences de l'Information (en considérant ici la pluridisciplinarité de ce domaine).
4. Langue du corpus Nous avons choisi la langue portugaise pour deux raisons :
5.
   i. afin d'acquérir des connaissances sur l'extraction des syntagmes nominaux dans cette langue et envisageant d'ores et déjà, dans le cadre d'une thèse de doctorat, la possibilité de développement d'un analyseur morpho-syntaxique pour cette langue ;
   ii. du fait que notre but initial était de travailler sur le développement de systèmes de recherche d'information sur des bases de données en langue portugaise. Ajoutons à ces raisons celle d'un engagement personnel avec l'institution qui nous a accordé notre bourse d'études.
   iii. Cependant, les résultats de ce travail, en ce qui concerne l'interface elle-même, pourront servir à des corpus dans d'autres langues, la maquette étant indépendante du traitement de l'information.

1. Niveaux des syntagmes nominaux Les syntagmes nominaux possèdent des relations d'emboîtement les uns par rapport aux autres. L'ordre de la relation d'emboîtement, appelé niveau, détermine la hauteur de l'arbre des syntagmes nominaux qui à son tour, restreint les possibilités de raffinement de la recherche d'information. Afin de construire l'arborescence permettant le raffinement d'une recherche d'information, nous avons choisi des articles ayant au moins des syntagmes nominaux de niveau quatre.

Une fois ces critères établis, nous avons choisi des articles publiés dans la revue brésilienne « Ciência da Informação », spécialisée en Sciences de l'Information. Cette revue est publiée et distribuée par l'Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). Les articles sélectionnés sont dans l'annexe A, dont les titres sont les suivants :

1. Conhecimento como recurso estratégico empresarialAnna Soledade VIEIRA. « Conhecimento como recurso estratégico empresarial ». Ciência da Informação. 1993, vol. 22, nº 2. p. 99-101. (La connaissance comme ressource stratégique des entreprises) - mots clés : ressources informationnelles ; intelligence compétitive ;
2. Inteligência competititva e decisão empresarialPatrick MAURY. « Inteligência competitiva e decisão empresarial ». Ciência da Informação. 1993, vol. 22, nº 2. p. 138-141. (L'intelligence compétitive et la prise de décision des entreprises) - mots clés : information ; intelligence compétitive ; gestion ; stratégies de décision ;
3. Economia da informação (L'économie de l'information)Pedro Onofre FERNANDES. « Economia da Informação ». Ciência da Informação. 1991, vol. 20, nº 2. p. 165-168. - mots clés : économie de l'information ; information / caractéristiques ; analyse du coût-bénéfice / coût / efficacité / performance / valeur ;
4. A Informação como insumo estratégicoDorodame Moura LEITÃO. « A informação como insumo estratégico ». Ciência da Informação. 1993, vol. 22, nº 2. p. 118-123. (L'information comme matière première stratégique) - mots clés : information stratégique ; systèmes d'information ; information opérationnelle ; gestion stratégique ;
5. Informação técnico-econômica: mais importante do que nuncaJoão Salvador FURTADO. « Informação técnico-econômica : mais importante do que nunca ». Ciência da Informação. 1991, vol. 20, nº 1. p. 20-22. (L'information téchnique-économique : plus important que jamais) - mots clés : information technologique ; information économique ; systèmes d'information technico-économique ; politique de recherche et de développement / entreprises ;
6. Perspectivas do agente da informação no contexto brasileiroDenise Werneck de PAIVA. « Perspectivas do agente da informação no contexto brasileiro ». Ciência da Informação.1990, vol. 19, nº 1. p. 48-52. (Perspectives de l'agent de l'information dans le contexte brésilien) - mots clés : agent de l'information ; bibliothécaire ; spécialiste de l'information ;
7. Sistemas de informação : a evolução dos enfoquesMarcos DANTAS. « Sistemas de

Informação : a evolução dos enfoques ». Ciência da Informação.1992, vol. 21, nº 3. p. 192-196. (Les systèmes d'information : l'évolution de ses approches) - mots clés : systèmes d'information ; théorie des systèmes ; services d'information ;

8. Consultoria informatológica em revisão : uma alternativa para serviços de informação personalizadosMariano A. MAURA. « Consultoria Informatológica em revisão : uma alternativa para serviços de informação personalizados ». Ciência da Informação.1993, vol. 22, nº 3. p. 242-247. (Consultation dans le domaine des sciences de l'information en révision : une alternative pour les services d'information personnalisés) - mots clés : services d'information ; bibliothèques spécialisées ; consultation dans le domaine des sciences de l'information ;

9. Informação para a indústriaMarisa Gurjão PINHEIRO. « Informação para a Indústria ». Ciência da Informação. 1991, vol. 20, nº 1. p. 16-19. (L'information pour l'industrie) - mots clés : information industrielle ; transfert de l'information ; information technologique ; information technologique / petite et moyenne industrie / Brésil ;

10. Interação entre empresas com necessidades de informação (=conhecimento) e a estrutura nacional de centros com provisão de conhecimento acumulado : referência especial à estrutura nacional de serviços de informação, documentação e de bibliotecaKjeld KLINTOE. « Interação entre empresas com necessidades de informação (=conhecimento) e a estrutura nacional de centros com provisão de conhecimento acumulado : referência especial à estrutura nacional de serviços de informação, documentação e de biblioteca ». Ciência da Informação. 1991, vol. 20, nº 1. p. 55-57. (Interaction entre les entreprises ayant besoin d'information (=connaissances) et la structure nationale de centres ayant un fonds : référence spéciale à la structure de services d'information, de documentation et de bibliothèques) - mots clés : politique d'information ; transfert d'information ; flux d'information ; centres et services d'information ; information technologique ;

11. Uso da informação na indústria como paradigma para o desenvolvimento econômicoFrancisco das Chagas de SOUZA. « Uso da informação na indústria como paradigma para o desenvolvimento econômico ». Ciência da Informação. 1991, vol. 20, nº 1. p. 34-36. (L'utilisation de l'information dans l'industrie comme paradigme pour le développement économique) - mots clés : information / développement économique ; information technologique ; information industrielle ; information économique ; services d'information / entreprise ;

12. A Informação eficaz na empresaAuta Rojas BARRETO. « A informação eficaz na empresa ». Ciência da Informação. 1991, vol. 20, nº 1. p. 78-81. (L'information efficace dans l'entreprise) - mots clés : information technologique ; prospection technologique ; services d'information ; entreprise de consultation ; entraînement de gestion ; ressources humaines ; produits d'information ;

13. Gerência da informação: mudanças nos perfis profissionaisRegina de Barros CIANCONI. « Gerência da informação : mudança nos perfis profissionais ». Ciência da Informação. 1991, vol. 20, nº 2. p. 204-208. (La gestion de l'information : changement dans les profils professionnels) - mots clés : administration des ressources d'information ; gestion de l'information ; professionnel de l'information ;

Informação: instrumento de dominação e de submissãoVânia Maria Rodrigues de ARAÚJO. « Informação: instrumento de dominação e de submissão ».Ciência da Informação. 1991, vol. 20, nº 1. p. 37-43. (L'information : outil de domination et de soumission) - mots clés : transfert d'information ; information technologique ; politique d'information ; développement technologique ; politique de science et technologie ; transfert de technologie ; 14.

Informação: a chave para a qualidade totalVirgínia Bentes PINTO. « Informação : a chave para a qualidade total ». Ciência da Informação. 1993, vol. 22, nº 2. p. 133-137. (L'information : la clé pour la qualité totale) - mots clés : qualité totale ; information pour la qualité ; unités d'information ; systèmes d'information. 15.

Remarque : les articles ayant été indexés dans la revue Ciência da Informação, les mots-clés ont été traduits.

2 Traitement préalable du corpus

2.1 Saisie des articles

Pour gagner du temps, dans la procédure d'extraction des syntagmes nominaux et du chargement de la base de données dans la maquette, il a fallu enregistrer les articles (documents) sur l'ordinateur et les préparer au préalable pour le traitement de l'information. Nous avons voulu travailler sur l'ordinateur à partir du moment où les articles ont été choisis.

Pour enregistrer les articles (documents) sur l'ordinateur, on a utilisé un *scanner* à main *(ScanMan Logitech)* pour la numérisation des textes et le logiciel Omnipage Direct pour la reconnaissance des caractères. Ce *scanner* a été choisi pour la bonne raison qu'il a été livré avec un logiciel d'OCR (Optical Character Recognition) - ou logiciel de reconnaissance optique des caractères - capable de reconnaître les caractères de la langue portugaise. En fait ce logiciel traite des textes en onze langues (allemand, anglais, danois, espagnol, français, italien, irlandais/gaélique, néerlandais, norvégien, portugais, suédois). L'autre raison déterminante de ce choix, à l'époque (début de l'année 1995), était le prix, moins cher que les *scanners* à plat. Ainsi, bien que l'on sache que des questions d'ergonomie et de précision se posent pour ce type de *scanner*, on l'a quand même choisi car il n'y avait que quinze articles à traiter.

Le temps nécessaire pour numériser chaque article a été d'environ trois heures en moyenne. Ce temps, un peu long, s'explique car la bonne utilisation de ce *scanner* dépend fondamentalement de la dextérité de la main de l'opérateur. En outre, la qualité de l'impression du document à numériser et le réglage du contraste du *scanner* comptent beaucoup pour la précision de la numérisation et de la reconnaissance des caractères d'un texte.

Les caractères non reconnus par le logiciel d'OCR ont été remplacés par le caractère « @ ». Cependant, d'autres caractères ont été reconnus d'une manière incorrecte. Nous avons rencontré les problèmes suivants :

1. quelques lettres accentuées sont souvent prises pour d'autres lettres en fonction de la proximité de l'accent sur la lettre. Exemple : la lettre « ó » a parfois été reconnue comme un « 6 », la lettre « í » comme la lettre « f » et quelquefois comme la lettre « r » ;
2. la lettre « r » proche de la lettre « n » a parfois été reconnue comme étant la lettre « m » ;
3. l'inverse de la situation 'b', plus haut, a été constaté lorsque le mot « information » apparaissait dans le texte et que l'OCR l'a reconnu comme étant « inforrnation ». Il semble reconnaître la lettre « m » comme étant les lettres « r » et « n » ;
4. la lettre « i » est parfois reconnue comme étant la lettre « l » ;

Cette expérience a montré que pour un travail professionnel, il faut plutôt utiliser un *scanner* à plat et un logiciel d'OCR capable de résoudre les problèmes orthographiques dus à la méconnaissance des caractères, car la correction automatique de l'orthographe peut conduire le logiciel à adopter des mots qui n'ont rien à voir avec les mots du texte. Il faut donc choisir un logiciel qui puisse proposer aux utilisateurs le choix du mot correct, c'est-à-dire, un logiciel avec un maximum d'interactivité avec l'utilisateur ou l'opérateur du *scanner*.

2.2 Préparation des fichiers des articles

Pour le traitement des articles, nous avons choisi le logiciel Word version 6.0a. La préparation des articles a consisté à énumérer chaque paragraphe dans chaque article du corpus. Cette procédure a été nécessaire afin d'identifier chaque syntagme nominal ; cette identification des syntagmes nominaux permettra à son tour de retrouver les articles d'où ils ont été extraits. Ainsi, les syntagmes nominaux ont été identifiés par le numéro d'article et le numéro d'ordre du paragraphe d'où ils ont été extraits.

Exemple : ***a informação; 1; 4*** qui veut dire que le syntagme ***a informação*** a été extrait de l'article nº 1, paragraphe nº 4

M. LE GUERN considère que l'identification des syntagmes nominaux devrait être réalisée par le numéro de l'article et par le numéro de la ligne. Certes, cette procédure est plus précise. Or, pour cette expérimentation, la numérotation adoptée, bien que moins précise, n'a guère compromis les résultats car les articles n'étant pas longs, des paragraphes entiers étaient presque toujours visibles sur l'écran. Pour une application professionnelle cependant, l'adoption d'une technique précise de manière à identifier les syntagmes nominaux est souhaitable, soit au moyen de la numérotation des lignes, soit d'une autre façon qui puisse les distinguer lorsque les articles sont présentés à l'écran.

3 Extraction des syntagmes nominaux

Pour extraire des syntagmes nominaux d'un corpus il fallait d'abord les identifier. Pour cela, on a utilisé une démarche logico-sémantique, étant donné qu'il n'y a pas d'analyseur morpho-syntaxique pour le portugais, ni de règles pour cette identification.

M. LE GUERN montre (Le *Français Moderne*, juin 1991) qu'un syntagme nominal est le résultat de la mise en œuvre de deux organisations logiques différentes. Si on prend comme exemple le syntagme nominal « o sistema de informação » afin de l'adapter à l'explication de M. LE GUERN, on verra que : le mot « sistema », en tant qu'élément du lexique de la langue, ne désigne aucun objet quel qu'il soit, mais uniquement un ensemble de propriétés, sans la prise en compte d'un univers donné. Il relève d'une logique intensionnelle, logique sans référentiel et sans classe, constitué de relations et de propriétés envisagées indépendamment de quelque objet que ce soit. Dans ce cas, le mot « sistema » trouvera différentes acceptions dans plusieurs domaines. Par contre le terme « sistema do Nutec » prend sa valeur dans un univers précis du discours. C'est un « sistema » qui appartient à une institution appelée Nutec. C'est un prédicat lié. C'est le basculement de la logique intensionnelle à la logique extensionnelle. C'est la mise en relation des mots et des choses. Lorsqu'un déterminant opère sur ce terme, « o sistema do Nutec » on a le syntagme nominal. Dans ce raisonnement, on identifie aussi le centre du syntagme, c'est-à-dire « sistema », qui est un prédicat libre.

3.1 Mise en forme des syntagmes nominaux

Afin d'éviter les problèmes de tri et par suite l'imprécision des résultats pendant la recherche d'information, nous avons adopté les dispositions suivantes pour la mise en forme des syntagmes nominaux :

1. saisie des syntagmes nominaux en caractères minuscules Les mots apparaissent dans les textes sous plusieurs formes, soit commençant avec une lettre majuscule, soit tout en minuscule ou encore tout en majuscules. Les systèmes de gestion de bases de données utilisent le code de chaque lettre pour trier les mots. En conséquence, les résultats sont influencés puisqu'on ne peut pas, si nécessaire, retrouver un mot qui existe dans la base, étant donné qu'il est écrit ou enregistré de différentes façons. C'est pourquoi, il a fallu décider de la conversion de tous les caractères des syntagmes nominaux en minuscules.
2. suppression des accents et des cédilles Cette mesure a été adoptée car il y avait quelques fautes d'orthographe dans les articles choisis. Ainsi on a trouvé les mêmes syntagmes nominaux sous deux formes dont la seule différence était l'accent ou la cédille. D'autre part, ces signes sont souvent une source d'erreurs lorsqu'un utilisateur saisit une recherche. De plus, les claviers n'utilisent guère une même norme de disposition des touches : il y a des claviers du type américain, français et autre.

3.2 Calculs des syntagmes nominaux

Plusieurs situations se présentent dans les textes où le repérage des syntagmes nominaux n'est pas toujours évident. Cela arrive soit parce qu'il y a des éléments anaphoriques, soit parce qu'il y a des ellipses, soit encore parce qu'il y a d'autres situations où les syntagmes nominaux se trouvent cachés ; il est possible d'autre part, de trouver des syntagmes nominaux qui ne portent pas d'information. Ainsi, il a fallu adopter

quelques règles afin d'extraire les syntagmes nominaux de façon homogène :

1. Syntagmes nominaux vides Par principe les articles (documents) sont composés de sections et de parties dont les titres ont été considérés comme étant des syntagmes nominaux. Or, on s'est vite rendu compte que plusieurs de ces syntagmes ne portaient pas d'information, concernant le sujet du document, comme par exemple : Conclusão (Conclusion), Objetivo (Objectif), Antecedentes (Antécédents), Introdução (Introduction), etc. On les a alors supprimés de la liste des syntagmes nominaux. On a trouvé également dans les textes des syntagmes vides, tels que : nesse sentido (dans ce sens), nesse contexto (dans ce contexte), uma vez que (une fois que... ou étant donné que), tal processo (un tel processus...), outro angulo (sous un autre angle), o momento (à ce moment...), etc. Ces syntagmes ont été aussi supprimés.
2. Syntagmes nominaux cachés dans des phrases avec factorisation L'extraction des syntagmes nominaux dans des phrases avec factorisation n'est pas toujours évidente, sauf quand on a une indication claire du syntagme comme par exemple dans la phrase suivante : o processo de negociação dos setores privado e público Dans ce cas, le syntagme nominal de niveau 1 (un) est clairement distingué comme étant os setores privado e público, parce que le mot setoresest au pluriel et qu'il fait référence aux deux mots - privado e publico, au singulier - simultanément. Par contre, on a rencontré des situations où on a eu du mal à identifier le syntagme nominal de manière précise. Dans ces cas-là, on a décidé d'extraire le syntagme nominal composé par chaque mot de la suite coordonné et le complément de la phrase. Exemples :

3. i. Le syntagme nominal : a análise, interpretação, avaliação e comunicação da informação pelos meios convenientes a donné les syntagmes nominaux suivants :

- a análise da informação pelos meios convenientes
- a interpretação da informação pelos meios convenientes
- a avaliação da informação pelos meios convenientes
- a comunicação da informação pelos meios convenientes

1. Le syntagme nominal : o potencial de conhecimento e inteligência da organização a produit les syntagmes nominaux suivants :

- o potencial de conhecimento da organização
- o potencial de inteligência da organização
- Il existe également une autre forme de factorisation où les mots coordonnés apparaissent entre parenthèses comme un complément discriminatoire du terme ou de la phrase qui précède la parenthèse. Exemple : La construction : ***rapidez e profundas transformações (políticas, econômicas, sociais, tecnológicas)*** a produit les syntagmes nominaux :

- rapidez e profundas transformações políticas
- rapidez e profundas transformações econômicas
- rapidez e profundas transformações sociais
- rapidez e profundas transformações tecnológicas
- Cependant, il faut faire attention à la construction dans la parenthèse, parce qu'on ne peut adopter cette règle que pour les suites de mots coordonnés. Certes, il y a des constructions entre les parenthèses qui sont des phrases explicatives. Dans ces cas, les syntagmes nominaux qui apparaissent dans les parenthèses vont être extraits indépendamment. Or, si cette solution est apparemment facile lors d'une extraction manuelle, elle n'apparaît pas réalisable dans une procédure automatique.

1. Phrases entre guillemets Les guillemets sont utilisés normalement dans deux situations : soit pour distinguer une citation soit pour distinguer un mot ou un terme (groupe limité de mots). Dans le cas d'une citation, on a fait l'extraction des syntagmes nominaux comme dans un texte normal, tandis que pour le deuxième cas on a simplement enlevé les guillemets. On a trouvé, cependant, des situations où le terme dans les guillemets a été identifié comme un syntagme nominal. Exemple : a denominação de « Economia da Informação » Une fois, encore, on peut trouver des difficultés dans une procédure automatique d'extraction de syntagmes nominaux.
2. Phrase entre tirets Les phrases entre tirets ont été traitées de la même façon que les phrases entre parenthèses. La situation est similaire.
3. Déterminant Zéro À la différence du français, on trouve souvent en portugais des phrases où les articles sont omis, donc des phrases qui n'ont pas de déterminants. Selon M. ME GUERN, l'omission des articles indéfinis est plus courante dans les cas des substantifs abstraits au pluriel, comme par exemple : informações científicas, sistemas, etc. Celso CUNHA & Lindley CINTRACUNHA, C. et CINTRA, L. Nova Gramática do Português Contemporâneo. Lisboa : Edições João Sá da Costa, 1991. présentent également quelques situations où les articles sont omis :

- en cas d'énumération
- par accumulation Exemple : *perspectivas do agente da informação no contexto brasileiro: problemas, barreiras e desafios* dont on trouve les syntagmes nominaux suivants :
- perspectivas do agente da informação no contexto brasileiro
- o agente da informação no contexo brasileiro
- a informação
- o contexto brasileiro
- problemas
- barreiras
- desafios

- par dispersion Exemple : políticas, procedimentos, diretrizes e sistemáticas para a organização da função dont les syntagmes nominaux sont :
- políticas para a organização da função
- procedimentos para a organização da função
- diretrizes para a organização da função
- sistemáticas para a organização da função

- on peut supprimer l'article défini lorsque le substantif est abstrait ou dans un proverbe, ou dans des phrases de comparaison brève. *Exemple* : *Conhecimento como recurso estratégico empresarial* dont les syntagmes nominaux sont :
- conhecimento como recurso estratégico empresarial
- recurso estratégico empresarial
- avant les mots qui indiquent des disciplines d'étude utilisées avec les verbes aprender, *estudar, cursar, ensinar, et synonymes*. Exemple : *Aprender **Francês** , Cursar **Direito*** dont les syntagmes nominaux sont :
- francês
- direito

Calculs des anaphores Les éléments anaphoriques, en portugais, apparaissent souvent au moyen des particules suivantes : pronoms possessifs, pronoms démonstratifs, pronoms personnels, etc. L'extraction des syntagmes nominaux cachés par les éléments anaphoriques n'a pas toujours été facile. Lorsque les sources de ces éléments étaient près d'eux, on a pu les résoudre facilement. Par contre, quand leurs sources se situaient dans les paragraphes précédents ou encore plus loin l'extraction des syntagmes nominaux devenait très difficile. Malgré les difficultés rencontrées, nous avons cependant essayé de les résoudre. Deux cas d'anaphores cependant n'ont pas pu être résolus : d'une part les anaphores sans sources, tels que : nesse sentido (dans quel sens ? il n'y a pas de source dans le texte), desse modo (de quelle façon ? il n'y a pas de source dans le texte), nossa experiência (quelle expérience ? celle de l'auteur ? celle de techniciens d'information ?), etc. Et pourtant, il a été facile de constater que ces syntagmes ne portent aucune information, et qu'ils sont plutôt des termes accessoires dans le processus d'écriture. Le deuxième cas d'anaphore non résolu est celui des anaphores sans sources explicites, mais qui portent des informations du genre : esse período pré-industrial (ce période pré-industriel), esse sistema de comunicação (ce système de communication), aqueles benefícios que não podem ser mensurados monetáriamente,(ces bénéfices qui ne peuvent pas être mesurés financièrement), etc. Dans ces cas, les syntagmes ont été conservés et transcrits tels quels, sans aucun traitement. Bien qu'on ait résolu dans la plupart des cas les problèmes d'anaphores, les phrases obtenues sont parfois curieuses, comme par exemple : uma categoria de clientes conscientizados dos seus direitos a produtos e serviços 1.

de alta qualidade (une catégorie de clients conscientisés sur leurs droits à des produits et à des services de haute qualité) dont la solution est : □uma categoria de clientes conscientizados dos direitos dos clientes conscientizados a produtos e serviços de alta qualidade (une catégorie de clients conscientisés des droits des clients conscientisés à produits et services de haute qualité) Une manière de résoudre ce problème serait de remplacer les éléments anaphoriques seulement au moment de l'extraction des syntagmes qui les enveloppent. Ainsi, l'exemple ci-dessus reste :

SN4
uma categoria de clientes conscientizados dos seus direitos a produtos e serviços de alta qualidade(une catégorie de clients conscientisés sur leurs droits à des produits et à des services de haute qualité)

SN3
clientes conscientizados dos seus direitos a produtos e serviços de alta qualidade (des clients conscientisés sur leurs droits à des produits et à des services de haute qualité)

SN2
os direitos dos clientes conscientizados a produtos e serviços de alta qualidade (les droits des clients conscientisés à des produits et à des services de haute qualité)

SN1
produtos e serviços de alta qualidade (des produits et des services de haute qualité)Cette solution n'a pas été adoptée dans ce travail puisqu'on a préféré garder les syntagmes nominaux entièrement développés.

Calculs des ellipses Le problème lié à ce type de figure est toujours dépendant de la capacité de se rendre compte qu'il manque un mot dans une phrase. Il faut toujours analyser non seulement les phrases précédentes, mais aussi les phrases suivantes. Exemple : □uma visão de longo prazo que assegure não só a sobrevivência (?), como também o crescimento da organização (une vision à long terme qui assure non seulement la survie mais aussi la croissance de l'organisation) Quel est le complément du terme sobrevivência (survie), c'est-à-dire, la survie de qui ? La solution se trouve dans la phrase suivante : o crescimento da organização (la croissance de l'organisation). Ainsi, le syntagme complet est : □uma visão de longo prazo que assegure não só a sobrevivência da organização, como também o crescimento da organização (une vision à long terme qui assure non seulement la survie de l'organisation mais aussi la croissance de l'organisation). Dans une procédure manuelle il n'y a pas de problèmes pour trouver la solution des ellipses ; par contre, dans une procédure automatique on rencontrera sûrement des difficultés pour résoudre ce type de figure. 1.

4 Conclusion

Pendant la procédure d'extraction des syntagmes nominaux du corpus, bien qu'on ait pris toutes les précautions, il a été difficile de garder une certaine homogénéité. Les raisons principales sont : (a) la diversité de style de rédaction des articles, étant donné qu'ils ont été écrits par quinze auteurs différents. On s'est rendu compte qu'il y a un lien étroit entre la facilité d'extraction des syntagmes nominaux et la clarté des articles ; (b) le manque d'un ensemble de règles pour orienter plus précisément la procédure d'extraction des syntagmes nominaux ; (c) le processus d'extraction n'a pas été continu, car des événements divers ont interrompu le travail. Ainsi, il a fallu faire une révision des syntagmes nominaux extraits, au fur et à mesure du chargement de la base de données et de la construction de l'arborescence.

Les résultats de cette étape sont consolidés dans la figure 3.1, où on trouve le nombre de syntagmes nominaux avec et sans doublons, le nombre de mots, de paragraphes, de lignes et de pages dans chaque article.

Le nombre de syntagmes nominaux avec doublons tient compte de la multiplicité d'occurrence d'un même syntagme nominal dans un article donné, tandis que le nombre de syntagmes nominaux sans doublons ne le fait pas. On a constaté par ailleurs qu'il y a des doublons qui peuvent apparaître ou non sur l'ensemble des syntagmes nominaux appartenant à plus d'un article différent. La colonne des syntagmes nominaux sans doublons ne tient pas compte de cet aspect là.

Article	Nombre de Syntagmes		Nombre de			
	avec doublons	sans doublons	mots	paragraphes	lignes	pages
1	681	520	1 904	38	169	4
2	576	454	1 759	49	196	4
3	555	342	1 834	80	245	5
4	502	344	2 010	55	236	5
5	726	550	1 857	77	234	5
6	388	123	1 950	30	150	1
7	684	475	3 176	28	197	1
8	642	454	2 034	45	218	4
9	588	177	1 818	81	179	1
10	479	353	1 560	32	154	3
11	467	334	1 709	33	167	3
12	611	542	2 057	80	257	5
13	519	1 n	1 498	72	178	1
14	672	439	2 330	42	235	5
15	560	128	1 939	33	178	1
Total	8 818	6 635	28 937	770	3 059	69

A partir de ce tableau et en tenant compte du nombre de syntagmes nominaux avec doublons (cette variable a été prise parce qu'elle représente la totalité de syntagmes nominaux d'un article), on arrive aux moyennes suivantes :

- Nombre moyen de syntagmes nominaux par lignes = 2,88
- Nombre moyen de syntagmes nominaux par paragraphe = 12,60
- Nombre moyen de syntagmes nominaux par page [49] = 139,96
- Nombre moyen de mots par syntagmes = 3,21

On a utilisé la moyenne bien que l'on sache qu'il s'agit d'une mesure d'utilité douteuse pouvant être facilement influencée par une valeur trop grande (ou trop petite) lui faisant perdre complètement sa représentativité. Le but de la présentation du tableau des moyennes n'est pas de chercher une relation entre les variables, mais de décrire le travail d'extraction des syntagmes nominaux. Concernant les moyennes il faut tenir compte des aspects suivants :

[49] Les données descriptives des articles ont été pris dans des fichiers dont les caractères sont en Times New Roman, taille 11.

1. La taille des paragraphes varie largement. On trouve des paragraphes avec une seule ligne et d'autres avec dix lignes ou plus. Cet aspect dépend du style de rédaction des auteurs ;
2. En ce qui concerne la taille des pages, on peut dire qu'elle présente presque le même problème, étant donné que les dernières pages de chaque article ne sont pas toujours pleines ;
3. A propos du nombre moyen de syntagmes nominaux par ligne, bien que l'on trouve des lignes remplies à moitié, elles apparaissent en générale complètes ;
4. Au sujet du nombre moyen de mots par syntagmes nominaux, on peut croire qu'il représente plutôt les syntagmes de premier et de deuxième niveau et à la rigueur ceux de troisième niveau, mais en aucun cas il ne s'agirait des syntagmes de quatrième et de cinquième niveau. Ces derniers sont très longs, composés quelquefois de vingt mots ou plus. Ainsi, pour obtenir une valeur plus fiable il aurait fallu faire ce calcul plutôt à chaque niveau.

Pour une analyse plus approfondie et si l'on envisage de trouver des relations entre les variables présentées dans la figure 3.1, un travail spécifique mérite d'être faite en considérant des critères plus orientés vers ce type d'étude.

En ce qui concerne la procédure d'extraction des syntagmes nominaux, nous avons découvert quelques points assez importants qui se prêtent à des études plus approfondies. Ce sont :

Bien qu'on ait adopté des solutions pour chacun de ces points, l'analyse et la formalisation de solutions définitives sont indispensables.

« Il n'y a pas de grandeur pour qui veut grandir. Il n'y a pas de modèle pour qui cherche ce qu'il n'a jamais vu. » Eluard (Eugène Grindel, dit Paul), L'Évidence poétique (Gallimard).

Chapitre 4 Développement de la maquette d'un SRI

1 Considérations préliminaires

Tout d'abord, pour la construction d'une maquette d'un système de recherche d'information, il a fallu connaître quelles étaient les caractéristiques de l'interface de ce système et comment il devrait fonctionner. Dans les chapitres précédents nous avons déjà indiqué quelques caractéristiques souhaitables, comme celles-ci : l'interactivité ; éviter l'utilisation d'un langage de commande trop compliqué ; l'utilisation des outils graphiques.

Il y a deux sortes d'interface que l'on pourrait développer : 1) soit en utilisant le langage naturel pour maintenir une interaction libre entre le système et l'utilisateur ; 2) soit

au moyen d'une interaction en utilisant une interface guidée par menus créés dynamiquement, où le système offre tout d'abord aux usagers la possibilité de recherche et ensuite les aide à la raffiner. Il est vrai que d'autres types d'interface pourraient être adopter, mais on a vu les inconvénients de ces interfaces dans le chapitre premier. La première solution a l'avantage d'utiliser la langue parlée, mais elle oblige les usagers à saisir leurs demandes. Ce qui peut créer des ennuis comme les fautes d'orthographe et/ou des saisies parfois longues. Ce qui peut rendre l'interface peu conviviale. La deuxième procédure, qui passe par l'utilisation des menus, laisse aux usagers la possibilité de choisir les syntagmes pertinents, de demander au système la présentation des documents trouvés, ou alors de recommencer la recherche, tout en utilisant des outils graphiques et un périphérique de pointage. Ceci peut rendre l'interface plus conviviale et plus facile.

Acteur	Interaction				
Usager	(1) [illegible]				
[illegible]	(2) [illegible]				
Usager	(3) L'usager [illegible]				
[illegible]	(4) [illegible]				
	(5) Bouton I [illegible]	(6) Bouton II - Montrer les Documents [illegible]	(7) Bouton III - [illegible]		
Usager	(8) [illegible]	(9) [illegible]	(10) [illegible] FIN		
[illegible]	(11) [illegible]				
	(12) Bouton IV [illegible]	(13) Bouton V - [illegible]	(14) Bouton VI - [illegible]	(15) Bouton VII – [illegible]	(16) [illegible] FIN
Usager	(17) Pour raffiner la recherche [illegible]	(18) [illegible]	(19) Ce bouton permet à [illegible]	(20) Si l'usager veut [illegible]	.

Suivant la suggestion de M. LE GUERN, nous avons travaillé sur le développement d'une interface où la procédure de recherche d'information découlait d'une interaction étroite entre le système et l'usager. Ce qu'il montre dans la citation suivante :

« ... L'outil d'aide à l'interrogation donne, pour chaque syntagme d'un niveau donné, la liste des syntagmes de rang immédiatement supérieur qui le contiennent, avec le nombre d'occurrences de chaque syntagme dans le corpus. L'utilisateur peut ainsi cheminer dans l'ensemble des descripteurs, en réduisant progressivement le nombre d'occurrences, jusqu'au moment où il obtiendra la quantité de références correspondant à l'ordre de grandeur qu'il souhaite. » [50]

En fait l'esquisse de l'interface de recherche d'information présentée dans le chapitre 2 a eu comme base cette citation de M. LE GUERN. Ainsi, suivant cette idée, nous avons développé une séquence d'interaction qui servira de base pour la construction de la maquette. Cette séquence d'interaction est montrée dans la figure 4.1.

2 Choix de l'approche de développement de la maquette

Le développement d'une maquette peut être élaboré en utilisant plusieurs outils. Ces outils ont été choisis selon ce que l'environnement de la maquette imposait comme caractéristiques souhaitables. Ces dernières prises en compte ont été : 1) la souplesse de construction d'une maquette plus conviviale et de manipulation facile ; 2) la flexibilité pour pouvoir effectuer des modifications dans un court espace de temps ; 3) l'importation/exportation des données de/vers les fichiers en format Word 6.0a ; 4) la construction manuelle de l'arborescence ; 5) la disponibilité du logiciel.

Il existe fondamentalement deux grandes classes d'outils à analyser : 1) le développement de la maquette en utilisant la programmation conventionnelle, à partir de langages de programmation comme C++, Pascal, etc. ; 2) le développement de la maquette utilisant un logiciel de gestion de bases de données.

L'approche de développement de la maquette en employant des langages de programmation de haut niveau (C++, Pascal, etc.) possède l'avantage d'offrir les possibilités d'optimisation du temps de réponse et de construction de la maquette selon ce que l'on souhaite, rendant ainsi le système plus convivial et sur mesure. Or cette approche demande un peu plus de temps pour cette construction, étant donné qu'il faut programmer toutes les routines, les procédures d'accès aux structures de données, les procédures de contrôle et de formatage de l'écran, les procédures d'importation et d'exportation de fichiers, etc. Le temps dont nous avons disposé n'a pas été suffisant pour mettre en œuvre une telle maquette. En plus, la mise en place des changements de manière rapide est pratiquement impossible dans les conditions existantes.

L'approche de l'utilisation d'un système de gestion de bases de données, offre toutes les possibilités de création d'une maquette conviviale et de mise en place des ajustements nécessaires dans un court espace de temps. La compatibilité relative à l'échange de données (importation et exportation de données) avec le format Word 6.0a est en liaison avec le logiciel choisi. Il y a aujourd'hui plusieurs logiciels ayant une telle caractéristique. D'autre part, les applications développées sur les systèmes de gestion de bases de données ne sont pas toujours performantes puisqu'elles sont trop génériques et

[50] Michel LE GUERN. « Un analyseur morpho-syntaxique pour l'indexation automatique », Le Français Moderne. Juin, 1991, t. LIX, n°. 1, p. 34

demandent de grosses ressources de l'ordinateur. C'est pourquoi il est nécessaire de choisir un système de gestion de base de données plus approprié au type d'application qu'on veut développer.

Parmi les possibilités d'utilisation des systèmes de gestion de bases de données, on distingue principalement les systèmes de gestion de bases de données relationnelles et les systèmes de gestion de base de données textuelles. En principe les systèmes de gestion de bases de données textuelles (MINISIS, MicroIsis, Adhoc plus, Basis plus, etc.) sont plus adaptés au présent travail puisqu'ils possèdent des fonctions et des caractéristiques appropriés au traitement et à la recherche de l'information textuelle. Cependant, les interfaces de recherche et la procédure d'indexation sont déjà prêtes et n'admettent guère d'adaptation. C'est la raison pour laquelle nous n'avons pas pu travailler avec ces logiciels pour le développement de la maquette.

Ainsi, il ne restait que les systèmes de gestion de bases de données relationnelles, qui malgré la faible performance et le fait qu'ils ne soient pas appropriés au traitement de textes, permettent le développement des applications selon les caractéristiques souhaitables.

Le logiciel Access (construit par Microsoft) a été choisi en considérant les critères déjà discutés en plus des facteurs suivants :

Convivialité et souplesse de développement d'application 1.

- à cause de son interactivité et de son interface graphique, intégré à Windows, ce logiciel possède des outils d'aide et d'assistance à ceux qui en développent les applications.
- ces applications sont développées au moyen de la définition des paramètres, des macros, des procédures et d'un langage capable de gérer des procédures de type événementiel, etc.

Compatibilité avec les logiciels Word 6.0a et Excel. 1.

- le logiciel Access possède une grande facilité pour importer et exporter des données, surtout avec les logiciels Word 6.0a et Excel entre autres formats.

Rapidité de développement et de mise à jour d'application 1.

- facteur le plus important et déterminant pour le choix de ce logiciel, étant donné la limitation du temps pour ce travail. Cette facilité permet de changer au fur et à mesure des besoins les applications de manière assez rapide.

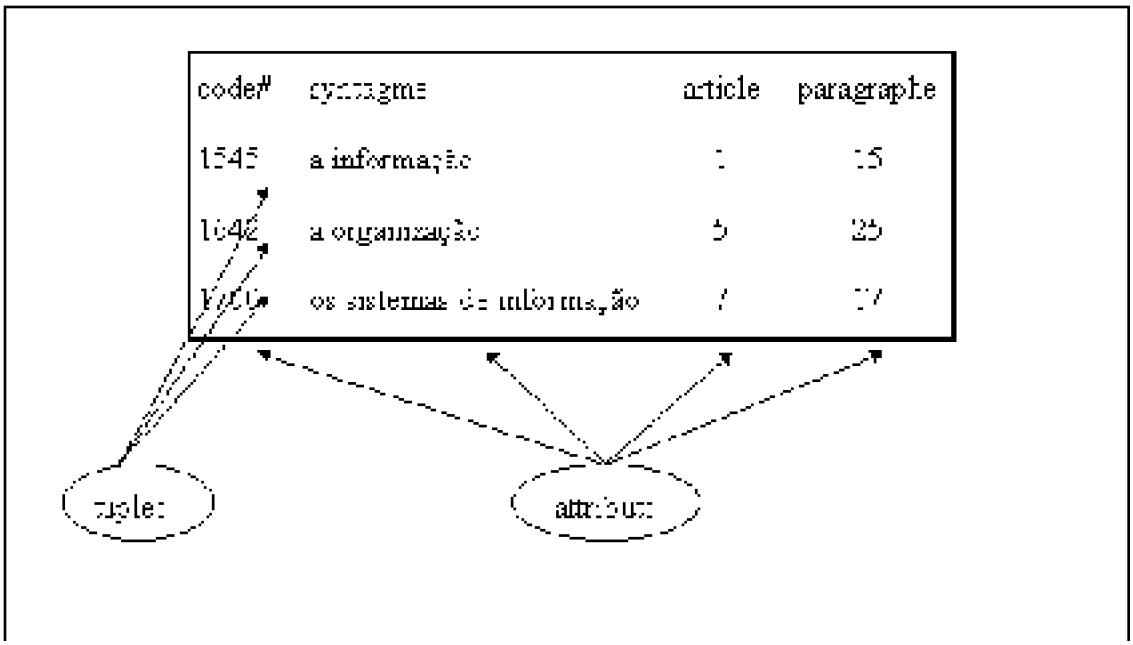
Tous ces éléments positifs ont été constatés pendant l'utilisation de ce logiciel. Il faut souligner en plus qu'il s'agit d'un logiciel d'apprentissage facile. La principale contrainte par contre, est la taille maximale d'un champ de 256 caractères. En conséquence, d'autres contraintes se produisent, parmi lesquelles l'impossibilité d'opérer des liaisons entre des champs ayant une telle taille.

3 Modèle de données relationnel

Le logiciel Access utilise l'approche des structures de données de type relationnel pour gérer les bases de données. Cette approche regroupe un ensemble de concepts tous déjà bien connus, mais qui méritent d'être présentés une fois de plus. Les définitions ont été empruntées à l'ouvrage de Max VETTER, *Modélisation des données : approches globales et orientée objets* [51] .

1. Tuple Une tuple est une liste de valeurs; une même valeur peut apparaître plusieurs fois. Exemple : <5, a informação, 7, 8> L'exemple ci-dessus signifie que : le syntagme nominal a informação, dont le code est 5, a été extrait de l'article 7, paragraphe 8.

2. Relation Une relation est un ensemble de tuples. Une relation est normalement représentée par une table où les colonnes désignent les attributs - chaque colonne contient les valeurs d'un unique et seul domaine - et les lignes forment les tuples. Une certaine tuple ne peut apparaître qu'une seule fois dans la relation. Exemple :



1. Les caractéristiques d'une relation :

- elle a un nom ;
- elle contient de 0 à « n » tuples (lignes) ;
- elle contient de 1 (une) à « m » colonnes aussi appelées attributs ;
- à l'intérieur d'une relation tous les attributs ont un nom unique ;
- les valeurs d'un attribut sont issues du même domaine, c'est-à-dire les valeurs d'un

[51] Max VETTER. *Modélisation des données : Approches globale et orientée objets*. Paris : Dunod Informatique, 1992

attribut ont les mêmes caractéristiques (numérique, alphabétique ou alphanumérique) ;

- une relation a au moins une clé ; la clé est un attribut ou un ensemble minimal d'attributs capables d'identifier de façon univoque les tuples de la relation ;
- chaque relation possède une clé primaire ;
- parfois des situations se présentent où il est nécessaire d'utiliser des clés composées de plus d'un attribut ; c'est le cas, par exemple, des syntagmes nominaux avec double rection, où un même syntagme nominal est lié avec deux syntagmes de niveau 1 distincts. Ainsi, la clé doit être formée par le syntagme de niveau 2 et celui de niveau 1, ce qui permet d'accéder au syntagme de deuxième niveau avec double rection, soit par un syntagme de niveau 1, soit par l'autre ; Dans la relation de la figure 4.3, on voit le besoin de créer une clé composée par deux attributs puisqu'un même syntagme de niveau 2 peut être lié à plusieurs syntagmes de niveau 1.
- une relation peut être représentée de la manière suivante : R(code#, syntagme, article, paragraphe)

Syntagme de Niveau 2	Syntagme de Niveau 1
a avaliação da informação pela empresa	a informação
a avaliação da informação pela empresa	a empresa
l'évaluation de l'information par l'entreprise	l'information
l'évaluation de l'information par l'entreprise	l'entreprise

Normalisation des relations Les relations non-normalisées présentent en général des 1. données redondantes. Cette redondance provoque sur le système une plus grande occupation de la mémoire, en plus du risque d'anomalies de mémorisation. Ces anomalies sont des difficultés liées aux opérations d'insertion, de modification et de suppression, qui peuvent amener les relations vers un état ne correspondant pas à une description de la réalité. Pour corriger ces problèmes, E.F. CODDCODD, E. F. « A relational model for large shared data banks ». CACM. 1970, vol. 13, nº 6.,CODD, E. F. « Further normalization of the relational model ». Data Base Systems, Courant computer science symposium 6, 1971. Rustin R. Editeur, Englewood Cliffs, New Jersey 1972. a été le premier à appliquer les règles de normalisation, au nombre de cinq aujourd'hui. Lorsqu'une relation respecte toutes ces cinq règles, on l'appelle une relation entièrement normalisée. Cependant, dans la pratique, on se contente des relations qui respectent les trois premières règles, celles qu'on appelle les relations

en troisième forme normale (ces règles sont connues aussi comme 1FN - Première Forme Normale, 2FN - Deuxième Forme Normale, 3FN - Troisième Forme Normale, 4FN - Quatrième Forme Normale, 5FN - Cinquième Forme Normale).

2. i. Première Forme Normale Dans une relation en première forme normale, les attributs ne peuvent prendre que des valeurs simples. Ainsi, à l'intersection d'une colonne (attribut) et d'une ligne (tuple) il ne peut y avoir qu'une valeur. Autre forme de définition de cette règle : une relation est en première forme normale si tous les attributs non-clés sont dépendants fonctionnels de la clé. Un attribut est dit dépendant fonctionnel d'une clé si à chaque valeur d'une clé ne correspond qu'une seule valeur de l'attribut.

ii. Deuxième Forme Normale Une relation en deuxième forme normale est caractérisée par le fait que tous ses attributs non-clés dépendent fonctionnellement de toute la clé (critère de 1FN) et non seulement d'une partie de la clé. C'est-à-dire qu'une relation est en 2FN si elle respecte la 1FN et que chaque attribut non-clé est entièrement dépendant fonctionnel de la clé. Cette deuxième forme normale est due au fait qu'une clé peut être formée par plus d'un attribut. Ce critère s'applique aux relations dont la clé est composée d'un minimum de deux attributs ayant au moins un attribut non-clé.

iii. Troisième Forme Normale Une relation est en troisième forme normale si elle respecte la 2FN et qu'elle ne porte aucune dépendance transitive (c'est-à-dire, il ne doit pas y avoir de dépendance fonctionnelle entre des attributs qui ne sont pas des clés candidates). Une relation peut avoir d'autres clés en plus de la clé primaire, celles-ci étant des clés candidates, lesquelles à leur tour sont en liaison simple les unes avec les autres.

Voici les principaux concepts pour la modélisation relationnelle de données. Il ne semble pas nécessaire de discuter plus en détail le modèle relationnel dans ce travail, car ce qui compte c'est la bonne utilisation de ce modèle. On présentera ainsi, par la suite, le modèle de données des syntagmes nominaux pour la construction de la maquette.

4 Modèle de données pour les syntagmes nominaux

Pour la modélisation de données, il est nécessaire tout d'abord de connaître le contexte du corpus et les faits qui caractérisent les syntagmes nominaux. Il faut également tenir compte des considérations présentées dans la section 1 de ce chapitre du fait qu'elles correspondent non seulement à l'approche de l'interface de recherche d'information, mais aussi à la démarche interactive entre l'utilisateur et la maquette de recherche d'information.

À la lumière du travail de construction de la base de données, on distingue l'existence de deux entités. Selon VETTER :

« Une entité est un exemplaire différentiable et identifiable d'une chose, d'une personne ou d'un concept concret ou abstrait, pour lequel on doit gérer des

informations significatives. Il y a des auteurs qui considèrent qu'une association (par exemple, l'union d'une femme et d'un homme est aussi une entité). » [52]

Ainsi, les deux entités sont :

1. ARTICLE (document du corpus) ;
2. SYNTAGME NOMINAL.

Les articles et les syntagmes nominaux constituent ensemble un contexte bien défini ; pour la modélisation des données on a considéré les faits suivants :

1. Les articles sont numérotés en ordre séquentiel à partir de 1 ;
2. Pour chaque article les paragraphes sont aussi énumérés en ordre séquentiel à partir de 1 ;
3. A chaque article correspond un titre ;
4. A chaque article correspond un texte d'une longueur plus grande que 256 caractères ;
5. Un même syntagme nominal peut apparaître dans plusieurs articles ;
6. Un même syntagme nominal peut apparaître dans plusieurs paragraphes à l'intérieur d'un article donné ;
7. Les syntagmes nominaux peuvent être classés en cinq niveaux, selon le contexte de ce travail ;
8. Un même syntagme nominal peut être classé dans plus d'un niveau ;
9. Il doit exister une association entre les syntagmes nominaux d'un niveau donné avec ceux d'un niveau immédiatement inférieur (construction de l'arborescence) ;
10. Un syntagme nominal peut être associé à plusieurs syntagmes nominaux de niveau immédiatement supérieur ;
11. Plusieurs syntagmes nominaux peuvent être associés à un même syntagme nominal (le cas de double rection) ; ce syntagme à son tour peut appartenir à des niveaux distincts, ce qui dépend du niveau du syntagme nominal immédiatement inférieur ;
12. Il y a un centre de syntagme nominal associé à chaque syntagme nominal de premier niveau ;
13. Les syntagmes nominaux de premier niveau ont comme association de niveau inférieur les centres des syntagmes ;
14. Les mots associés aux syntagmes nominaux en dehors de l'ensemble des centres de syntagmes nominaux, sont aussi associés aux syntagmes de premier niveau en fonction de leur importance dans la recherche d'information. Par exemple :
15. i. Dans le cas « les systèmes d'information » — le centre du syntagme nominal est systèmes. Pourtant, le mot information est aussi important dans ce contexte que le mot systèmes.
    ii. Dans le cas « l'analyse d'information » — de même que dans l'exemple (a), on

[52] ***VETTER, Max. Modélisation des données : Approches globale et orientée objets. Paris : Dunod Informatique, 1992.***

considère analyse et information comme étant centres du syntagme nominal. Le premier « analyse » est le centre du syntagme nominal, le deuxième « information » c'est le centre complémentaire du syntagme nominal (voir dans le chapitre 5, à la section 4 'Centres complémentaires des syntagmes nominaux', la justificatif pour cette décision).

Les centres des syntagmes nominaux possèdent des flexions en genre et nombre. 1.

Définissons quelques termes qui seront utilisés dans la construction du modèle de données de façon à éviter des confusions. Ainsi :

on utilisera le terme TABLE au lieu de relation pour désigner un ensemble de tuples ; 1.

et le terme RELATION pour désigner une association entre deux TABLES ou plus. 2.

À partir des faits énumérés, on a conçu les structures de données nécessaires pour la construction de la maquette de recherche d'information. Toutes les tables conçues ont été soumises aux règles de normalisation. Elles sont ainsi en 3FN. Ces tables seront explicitées selon la nomenclature suivante :

- T (attr 1, attr 2, attr 3)
- où : T - est le nom de la table et sera représenté en majuscules
- attr 1 - l'attribut ou les attributs qui composent la clé ; il sera représenté en minuscules et sera souligné
- attr 2, attr 3... - sont les noms des attributs que appartiennent à la table ; ils seront représentés en minuscules

Les tables conçues sont donc :

ARTICLES (code-doc, titre, article) où : 1.

- code-doc : valeurs séquentielles (1 à 15) qui identifient l'article
- titre : contient le titre d'un article
- article : contient le texte d'un article Cette table est issue des faits 1, 3 et 4. Elle est en 3FN, parce que tous les attributs sont dépendants fonctionnellement de la clé et qu'il n'y a pas de dépendance transitoire entre les attributs non-clés.

SYNTAGMES (code du syntagme, syntagme) où : 1.

- code du syntagme : code numérique séquentiel qui identifie chaque syntagme
- syntagme - contient le syntagme nominal proprement dit
- Cette table, issue de la procédure de normalisation, est donc en 3FN.

SYNTAGMES NIVEAU 1 (code 1, syntagme 1, nombre d'articles) où : 1.

- code 1 : contient le code du syntagme nominal
- syntagme 1 : contient le syntagme de niveau 1
- nombre d'articles : contient l'information concernant le nombre d'articles du corpus où ce syntagme nominal apparaît.
- Cette table, en opposition à la table SYNTAGMES, ne contient que les syntagmes nominaux de niveau 1. Remarque : Cette table et les quatre autres suivantes sont redondantes par rapport à la table SYNTAGMES, mais elles sont toutes nécessaires car elles permettent une meilleure performance au sujet du temps d'accès étant donné que la sélection des syntagmes nominaux par niveau est faite avant la procédure de recherche. Elles sont toutes également issues du fait 7 ;

SYNTAGMES NIVEAU 2 (code 2, syntagme 2, nombre d'articles) où : 1.

- code 2 : contient le code du syntagme nominal
- syntagme 2 : contient le syntagme nominal de niveau 2
- nombre d'articles : contient l'information concernant le nombre d'articles du corpus où ce syntagme nominal apparaît.
- Cette table, en opposition à la table SYNTAGMES, ne contient que les syntagmes nominaux de niveau 2Voir remarque dans la description de la rubrique 'c' plus haut.;

SYNTAGMES NIVEAU 3 (code 3, syntagme 3, nombre d'articles) où : 1.

- code 3 : contient le code du syntagme nominal
- syntagme 3 : contient le syntagme nominal de niveau 3
- nombre d'articles : contient l'information concernant le nombre d'articles du corpus où ce syntagme nominal apparaît.
- Cette table, en opposition à la table SYNTAGMES, ne contient que les syntagmes nominaux de niveau 3Voir remarque dans la description de la rubrique 'c' plus haut. ;

SYNTAGMES NIVEAU 4 (code 4, syntagme 4, nombre d'articles) où : 1.

- code 4 : contient le code du syntagme nominal
- syntagme 4 : contient le syntagme nominal de niveau 4
- nombre d'articles : contient l'information concernant le nombre d'articles du corpus où ce syntagme nominal apparaît.
- Cette table, en opposition à la table SYNTAGMES, ne contient que les syntagmes nominaux de niveau 4Voir remarque dans la description de la rubrique 'c' plus haut. ;

SYNTAGMES NIVEAU 5 (code 5, syntagme 5, nombre d'articles) où : 1.

- code 5 : contient le code du syntagme nominal
- syntagme 5 : contient le syntagme nominal de niveau 5
- nombre d'articles : contient l'information concernant le nombre d'articles du corpus où ce syntagme nominal apparaît.
- Cette table, en opposition à la table SYNTAGMES, ne contient que les syntagmes nominaux de niveau 5Voir remarque dans la description de la rubrique 'c' plus haut. ;

CENTRE DU SYNTAGME (code du centre, centre du syntagme) où : 1.

- code du centre : code numérique séquentiel, identifie le centre du syntagme
- centre du syntagme : contient le mot qui joue le rôle de centre du syntagme nominal ou le mot qui joue le rôle de centre complémentaire du syntagme nominal (voir aussi dans le chapitre 5, section 4 la discussion sur le centre complémentaire de syntagme nominal) ;
- Cette table est issue des faits 12 et 14.

MOTS (code du centre, centre du syntagme) où : 1.

- code du centre : code du centre de syntagme nominal dont la valeur est égale à celle de la table CENTRE du SYNTAGME
- centre du syntagme : contient le mot résultant de la flexion en nombre d'un centre de syntagme nominal ou d'un mot qui équivaut à un centre de syntagme donné
- Cette table est issue du fait 15.

REFERENCE RESUMEE (code, article) où : 1.

- code : contient le code d'un syntagme nominal
- article : contient le code d'identification de l'article
- Cette table est conçue d'après le fait 5.

REFERENCE (code, article, paragraphe) où : 1.

- code : contient le code du syntagme nominal
- article : contient le code d'identification de l'article
- paragraphe : contient le code du paragraphe d'un article donné
- Cette table est issue des faits 2 et 6.

LIAISON CS - SN 1 (code syntagme niveau 1, code centre du syntagme) où : 1.

- code syntagme niveau 1 : contient le code d'un syntagme nominal de niveau 1

- code centre du syntagme : contient le code du centre de syntagme nominal
- Cette table est issue des faits 12 et 13.

LIAISON SN 1 - SN 2 (code syntagme niveau 2, code syntagme niveau 1) où : 1.

- code syntagme niveau 2 : contient le code du syntagme nominal niveau 2
- code syntagme niveau 1 : contient le code du syntagme nominal niveau 1
- Cette table est issue des faits 8, 9, 10 et 11.

LIAISON SN 2 - SN 3 (code syntagme niveau 3, code syntagme niveau 2) où : 1.

- code syntagme niveau 3 : contient le code du syntagme nominal niveau 3
- code syntagme niveau 2 : contient le code du syntagme nominal niveau 2
- Cette table est issue des faits 8, 9, 10 et 11.

LIAISON SN 3 - SN 4 (code syntagme niveau 4, code syntagme niveau 3) où : 1.

- code syntagme niveau 4 : contient le code du syntagme nominal niveau 4
- code syntagme niveau 3 : contient le code du syntagme nominal niveau 3
- Cette table est issue des faits 8, 9, 10 et 11.

LIAISON SN 4 - SN 5 (code syntagme niveau 5, code syntagme niveau 4) où : 1.

- code syntagme niveau 5 : contient le code du syntagme nominal niveau 5
- code syntagme niveau 4 : contient le code du syntagme nominal niveau 4
- Cette table est issue des faits 8, 9, 10 et 11.

TABLE GROS INDEX (code du syntagme, syntagme, article, paragraphe, niveau, centre du syntagme, syntagme niveau inférieur) où : 1.

- code du syntagme : est un code séquentiel numérique identifiant chaque occurrence des syntagmes nominaux, y inclus les syntagmes répétés.
- syntagme : contient le syntagme nominal
- article : contient le code de l'article d'où le syntagme a été extrait
- paragraphe : contient le code du paragraphe d'où le syntagme a été extrait
- niveau : contient le code du niveau du syntagme nominal
- centre du syntagme : indique si le contenu de l'attribut du syntagme nominal de niveau inférieur est un centre de syntagme ou non
- syntagme de niveau inférieur : contient le syntagme nominal de niveau inférieur par

rapport à l'attribut du syntagme

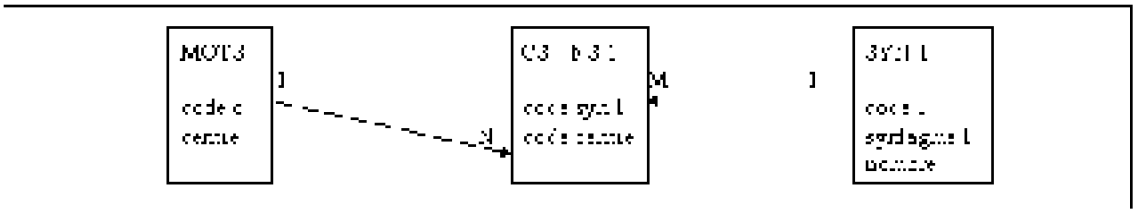
- La TABLE GROS INDEX est une table de travail d'où sont issues toutes les autres tables.

L'approche adoptée pour la construction de ce modèle de données a considéré l'utilisation du code identificateur d'un syntagme nominal au lieu d'utiliser le syntagme nominal lui-même comme étant la clé de chaque table. Cette option a été nécessaire à cause de la limitation du logiciel Access concernant la taille des champs d'une table : 256 caractères. En outre, la comparaison entre deux champs numériques est plus performante que la comparaison entre deux champs textuels. Ce problème est encore plus important lorsque des syntagmes nominaux de niveau 4 et 5 atteignent parfois la limite de 256 caractères. Ainsi, pour réussir le développement de la maquette, nous avons choisi l'utilisation des codes des syntagmes nominaux au lieu des syntagmes eux-mêmes, ce qui explique la création d'un nombre plus grand de tables.

5 Structure de données : navigation dans l'arbre des syntagmes nominaux

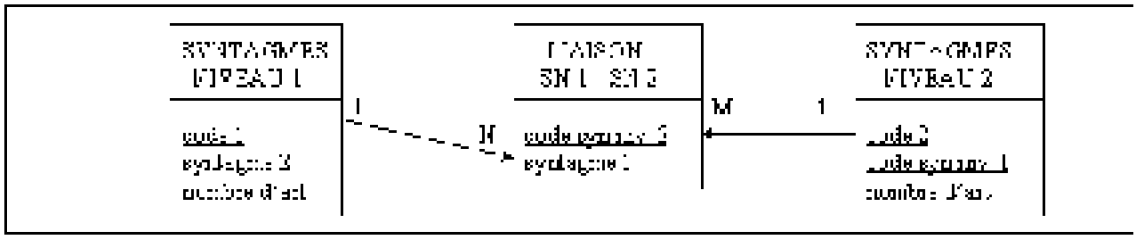
La structure de données permettant la navigation dans l'arborescence des syntagmes nominaux étant déjà prête, il reste à mettre en relation les tables de cette structure pour construire effectivement la navigation dans l'arborescence.

1. La recherche des syntagmes nominaux de niveau 1 (un) Cette recherche est faite à partir d'une demande de l'utilisateur qui consiste à saisir un mot. Le système cherche dans la table MOTS si le mot demandé existe, si oui, il détecte le code du centre du syntagme nominal correspondant. Dès qu'il a découvert ce code, le système trouve immédiatement dans la table LIAISON CS - NS 1 tous les codes des syntagmes de premier niveau associés à ce centre de syntagme. Les syntagmes nominaux vont être trouvés dans la table SYNTAGMES NIVEAU 1. De cette façon, on a conçu la relation suivante :



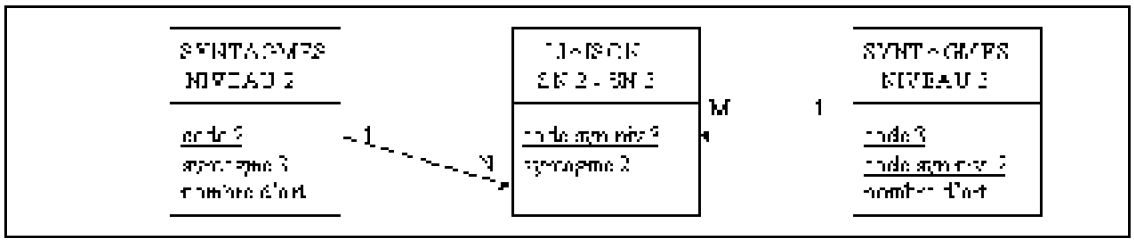
Dans la figure 4.4 on a abrégé les noms des attributs pour mieux placer le dessin, le nom complet de chaque attribut se trouvant dans la section 4 de ce chapitre. Il est important de noter que le schéma montre que l'on peut trouver, pour chaque centre de syntagme, plusieurs syntagmes de premier niveau. La figure montre aussi que l'inverse peut arriver, étant donné qu'un syntagme nominal de premier niveau peut avoir deux centres de syntagme nominal, selon la discussion dans le chapitre 5, section 4, « Centres complémentaires des syntagmes nominaux ».

1. La recherche des syntagmes nominaux de niveau 2 (deux) La recherche des syntagmes nominaux de niveau 2 (deux) est faite à partir du choix du syntagme nominal de premier niveau, demandé par l'utilisateur. Le système trouve le code du syntagme choisi dans la table SYNTAGMES NIVEAU 1 puis cherche dans la table LIAISON SN 1 - SN 2 tous les syntagmes de deuxième niveau qui sont associés au syntagme nominal du premier niveau choisi. Pour faire paraître les syntagmes associés, le système utilise la table SYNTAGMES NIVEAU 2 (voir figure 4.5).

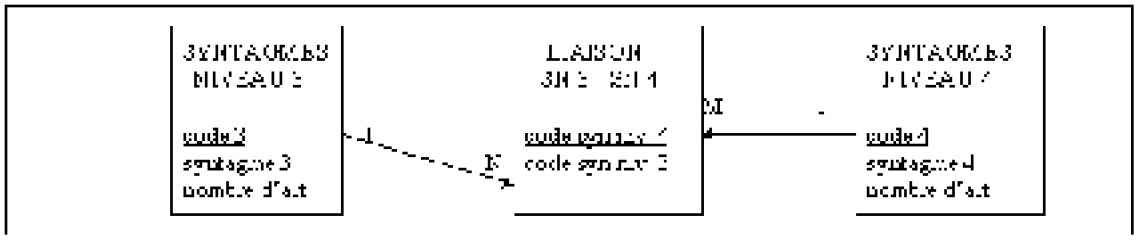


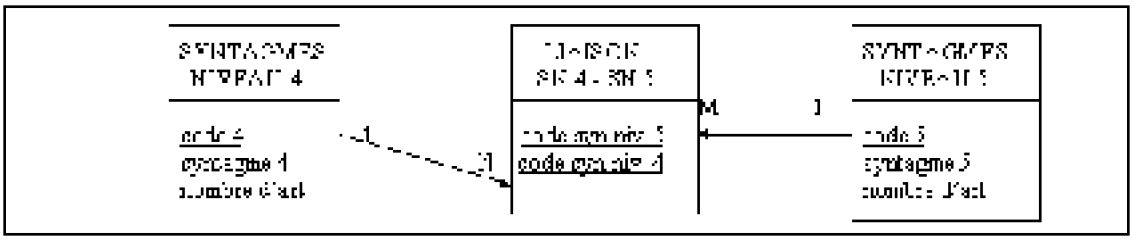
Ici on voit aussi que la structure LIAISON SN 1 - SN 2 permet qu'un syntagme quelconque de niveau un puisse être associé à plusieurs syntagmes de niveau deux et vice-versa.

1. La recherche des syntagmes de niveau trois La démarche pour cette recherche est la même que pour les recherches précédentes. Il suffit donc de changer les tables (voir la figure 4.6).

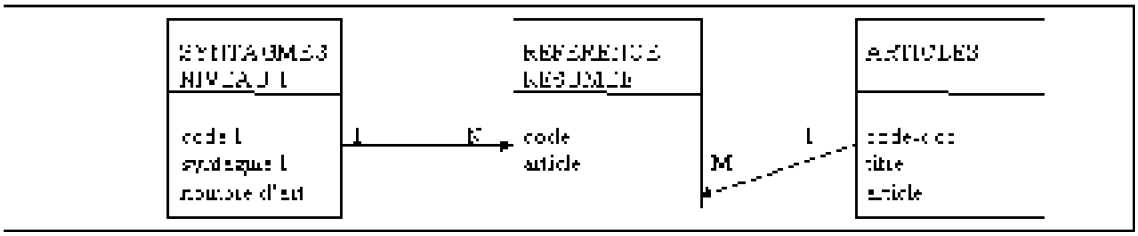


1. La recherche des syntagmes nominaux de niveau quatre et cinq La recherche des syntagmes nominaux de niveau quatre et cinq a été effectué de la même façon que toutes les autres déjà décrites (voir les figures 4.7 et 4.8 respectivement).





La recherche des titres des articles à partir d'un syntagme nominal choisi Pour chercher les titres qui correspondent aux articles d'où un syntagme nominal choisi a été extrait, on conçoit la relation suivante : 1.



On voit dans la figure 4.9 que pour un syntagme nominal de niveau 1 (un), le système prend le code du syntagme nominal dans sa table et cherche dans la table REFERENCE RESUMEE tous les articles d'où ce syntagme a été extrait. A partir du code de chaque article, le système accède au titre respectif. Cette démarche est analogue pour les syntagmes nominaux d'un autre niveau, il est suffisant d'opérer la modification de la table des SYNTAGMES NIVEAU 1 à une autre table qui corresponde au niveau du syntagme souhaité.

6 Développement de la maquette du Système de Recherche d'Information

Développer une maquette utilisant le logiciel Access signifie développer une application Access. Une telle application est composée d'objets d'utilisation directe comme les formulaires et les états, et d'objets d'utilisation indirecte comme les tables, les requêtes, les macros et les modules. Ces objets ont des propriétés paramétrables de façon à ce que l'on puisse donner l'aspect voulu à ces applications.

Par la suite on donnera une brève définition de ces objets, selon le manuel du logiciel.

« Un formulaire ... identifie les données à recueillir. Il permet d'enregistrer des informations dans une base de données, de les affranchir et les imprimer. » [53] ***« ... un état permet d'extraire et de présenter les données dans le format le mieux adapté à leur exploitation et à leur diffusion. Des étiquettes de publipostage, des factures, des rapports ... »*** [54] ***« Une table regroupe les données de même nature,***

[53] *MICROSOFT CO. Microsoft Acces : Guide de l'utilisateur. Ireland : 1994. p. 386*

[54] *Ibidem p. 592.*

... chaque enregistrement d'une table contient des informations sur un élément en particulier,... les enregistrements d'une table sont constitués de champs. » [55]
«Une requête sert à interroger des tables sur les données qu'elles contiennent. Sa structure indique précisément ... quelles données extraire. » [56]

Pendant l'utilisation d'un formulaire, la modification des données d'un champ, l'action de cliquer sur un bouton ou déplacer la souris sont identifiés comme des événements, auxquels le logiciel Access peut réagir automatiquement. Pour personnaliser cette réaction du logiciel, on peut utiliser les macros et les procédures événementielles. Ces macros ou procédures sont exécutés lorsqu'un événement donné se produit. Pour cela, il faut rattacher ces macros ou procédures à la propriété concernant l'événement. Un macro est composé d'un ensemble d'actions prédéfinies comme OuvrirFormulaire, DéplacerDimensioner, CopierObjet etc. Une procédure est écrite en langage Access Basic, composé par des commandes et des fonctions (équivalents aux actions des macros).

Une fois que les caractéristiques du logiciel Access ont été connues et que son fonctionnement a été compris, on s'est rendu compte que le développement de la maquette consistait uniquement en la création des **tables**, **requêtes** et **formulaires**. L'objet **état** n'a pas été utilisé parce qu'on n'a pas eu besoin de créer des fonctions d'impression de rapports.

6.1 Construction des Tables

Les tables ont été construites selon le modèle de données décrit dans la section 4 de ce chapitre. Liste des tables obtenues :

1. Articles Contient les codes, les titres et le contenu des articles.
2. Table des Syntagmes Contient tous les syntagmes nominaux indépendamment de leurs niveaux.
3. Table des Mots Contient les codes des centres de syntagmes et les flexions en nombre de ces syntagmes.
4. Table Référence Contient l'association entre les codes de chaque syntagme nominal, le code des articles et les paragraphes d'où ils ont été extraits.
5. Table Référence Résumée Contient l'association des codes des syntagmes nominaux et les codes des articles d'où ces syntagmes ont été extraits.
6. Table Centre du Syntagme Contient le code du centre de syntagme et les centres de syntagmes eux-mêmes.
7. Table Syntagme Niveau 1 Contient le code des syntagmes nominaux de premier niveau et les syntagmes nominaux eux-mêmes.
8. Table Syntagme Niveau 2 Contient le code des syntagmes nominaux de deuxième niveau et les syntagmes nominaux eux-mêmes.

[55] *Ibidem p. 136.*

[56] *Ibidem p. 240*

9. Table Syntagme Niveau 3 Contient le code des syntagmes nominaux de troisième niveau et les syntagmes nominaux eux-mêmes.
10. Table Syntagme Niveau 4 Contient le code des syntagmes nominaux de quatrième niveau et les syntagmes nominaux eux-mêmes.
11. Table Syntagme Niveau 5 Contient le code des syntagmes nominaux de cinquième niveau et les syntagmes nominaux eux-mêmes.
12. Table liaison CS - SN 1 Contient l'association entre les codes des syntagmes de premier niveau et les codes des centres des syntagmes auxquels ils sont liés
13. Table liaison SN 1 - SN 2 Contient l'association entre les codes des syntagmes de deuxième niveau et les codes des syntagmes de premier niveau, auxquels ils sont liés.
14. Table liaison SN 2 - SN 3 Contient l'association entre les codes des syntagmes de troisième niveau et les codes des syntagmes de deuxième niveau auxquels ils sont liés.
15. Table liaison SN 3 - SN 4 Contient l'association entre les codes des syntagmes de quatrième niveau et les codes des syntagmes de troisième niveau auxquels il est lié.
16. Table liaison SN 4 - SN 5 Contient l'association entre les codes des syntagmes de cinquième niveau et les codes des syntagmes de quatrième niveau auxquels ils sont liés.
17. Table Gros Index Contient les champs de code du syntagme, syntagme, article, paragraphe, niveau, centre du syntagme, syntagme niveau inférieur.

6.2 Construction des Requêtes

Pour la navigation dans l'arborescence des syntagmes nominaux nous avons construit des requêtes au moyen des relations présentées dans la section 5 de ce chapitre.

Les requêtes construites sont les suivantes :

1. Requête sur les SN 1 Cette requête cherche, à partir d'un code de centre de syntagme, tous les syntagmes de niveau 1 associés.
2. Requête sur les SN 2 Cette requête cherche, à partir d'un code de syntagme de premier niveau, tous les syntagmes de niveau 2 associés.
3. Requête sur les SN 3 Cette requête cherche, à partir d'un code de syntagme de deuxième niveau, tous les syntagmes de niveau 3 associés.
4. Requête sur les SN 4 Cette requête cherche, à partir d'un code de syntagme de troisième niveau, tous les syntagmes de niveau 4 associés.
5. Requête sur les SN 5 Cette requête cherche, à partir d'un code de syntagme de quatrième niveau, tous les syntagmes du niveau 5 associés.
6. Requête pour voir les titres 1 Cette requête cherche, à partir d'un code de syntagme de premier niveau, tous les titres des articles d'où il a été extrait.

Requête pour voir les titres 2 Cette requête cherche, à partir d'un code de syntagme 7. de deuxième niveau, tous les titres des articles d'où il a été extrait.

Requête pour voir les titres 3 Cette requête cherche, à partir d'un code de syntagme 8. de troisième niveau, tous les titres des articles d'où il a été extrait.

Requête pour voir les titres 4 Cette requête cherche, à partir d'un code de syntagme 9. de quatrième niveau, tous les titres des articles d'où il a été extrait.

Requête pour voir les titres 5 Cette requête cherche, à partir d'un code de syntagme 10. de cinquième niveau, tous les titres des articles d'où il a été extrait.

6.3 Construction des Formulaires

La construction des formulaires a été faite afin de procéder à l'interface entre l'utilisateur et l'application de la recherche d'information. Ainsi, les formulaires construits, dans l'ordre de présentation, sont les suivants :

Menu Général C'est le premier formulaire, dans l'ordre de présentation. Son objectif 1. est de présenter les options de l'application. L'utilisateur peut ainsi avoir le choix de la tâche à exécuter dans l'application. On a créé les boutons d'options suivants :

- Construction de l'arborescence des syntagmes nominaux Cette option permet de construire et d'ajuster l'arborescence des syntagmes nominaux. Elle permet aussi la création de nouveaux syntagmes nominaux. Une procédure événementielle associée a ce bouton ouvre le **Formulaire Saisir Syntagmes**.
- Recherche d'information Ce bouton associé à une procédure événementielle ouvre le formulaire **Formreq**. À partir de ce formulaire l'utilisateur est guidé dans l'arborescence des syntagmes nominaux.
- Base de données Cette option ouvre le module de création et de mise à jour de bases de données du logiciel Access. C'est une manière de mettre à jour la définition de la base de données et de l'application.
- Quitter l'application Ce bouton lorsqu'il est activé, permet de quitter l'application au moyen d'une procédure événementielle.

Formreq Une fois l'option de recherche d'information choisie, l'application ouvre le 1. formulaire Formreq. Ce formulaire a comme objectif principal de recevoir la demande des utilisateurs au moyen d'un mot ou d'un centre de syntagme nominal. En réponse, l'application cherche les syntagmes nominaux de premier niveau associés à cette demande et les présente, s'ils s'y trouvent. Ce formulaire a été développé en utilisant un sous-formulaire appelé syntagme de premier niveau ; ce sous-formulaire a comme fonction de montrer les syntagmes de premier niveau. La recherche de ces syntagmes est faite au moyen de la requête Requête sur les SN 1. Une fois présentés tous les syntagmes nominaux de premier niveau, les utilisateurs peuvent choisir une des options suivantes :

- Demander la recherche des syntagmes nominaux de deuxième niveau Cette option peut être activée dès que l'utilisateur a choisi un syntagme de premier niveau. L'activation de cette option est faite à partir d'une clique sur le bouton correspondant à cette option ; une procédure événementielle est alors exécutée pour chercher les syntagmes et ouvrir le formulaire **Voir syntagmes niveau deux** ;
- Voir les titres d'où le syntagme de premier niveau choisi a été extrait Une fois choisi un syntagme de premier niveau, l'utilisateur peut activer cette option ouvrant ainsi le formulaire **Voir les articles 1** au moyen d'une procédure événementielle ;
- Quitter l'option de recherche d'information De même que dans tous les autres cas ci-dessous, une fois le bouton concerné activé, l'application exécute une procédure événementielle qui ferme le formulaire **Formreq** et passe le contrôle au formulaire **Menu Général**.

Voir syntagmes niveau deux Ce formulaire est constitué d'un sous-formulaire, 1. syntagme de deuxième niveau, qui a pour fonction de présenter les syntagmes de niveau deux à partir du choix d'un syntagme de premier niveau et de la demande par l'utilisateur de rechercher ceux du niveau deux. Il utilise la requête Requête sur les SN 2. Le formulaire permet grâce à différents boutons d'option de :

- Chercher les syntagmes du troisième niveau La procédure associée à ce bouton, cherche les syntagmes de niveau trois et ouvre le formulaire **Voir syntagme niveau trois** ;
- Voir les titres une fois choisi un syntagme de niveau deux La procédure ouvre le formulaire **Voir les articles 2** ;
- Quitter le formulaire courant vers le formulaire **Formreq** Si cette option est choisie, l'application ferme le formulaire courant
- Faire une nouvelle recherche à partir d'un mot ou d'un centre de syntagme Si cette option est choisie, l'application ferme le formulaire courant
- Quitter l'application Si cette option est choisie, l'application ferme le formulaire courant et le formulaire **Formreq**.

Voir syntagmes niveau trois Ce formulaire est constitué d'un sous-formulaire, 1. syntagme de troisième niveau, qui a la fonction de présenter les syntagmes de niveau trois à partir du choix d'un syntagme de deuxième niveau et de la demande par l'utilisateur de rechercher ceux du niveau trois. Il utilise la requête Requête sur les SN 3. Le formulaire permet grâce à différents boutons d'option de :

- Chercher les syntagmes de quatrième niveau ; La procédure associée à ce bouton, cherche les syntagmes de niveau quatre et ouvre le formulaire **Voir syntagme niveau quatre** ;
- Voir les titres une fois choisi un syntagme de niveau trois La procédure ouvre le formulaire **Voir les articles 3** ;

- Quitter le formulaire courant vers le formulaire **Voir syntagme niveau deux** Si cette option est choisie, l'application ferme le formulaire courant ;
- Faire une nouvelle recherche à partir d'un mot ou d'un centre de syntagme Si cette option est choisie, l'application ferme le formulaire courant et le formulaire **Voir syntagmes niveau deux** ;
- Quitter l'application Si cette option est choisie, l'application ferme le formulaire courant, le formulaire **Voir syntagmes niveau deux** et le formulaire **Formreq**.

Voir syntagmes niveau quatre Ce formulaire est constitué d'un sous-formulaire, 1. Syntagmes de quatrième niveau, qui a la fonction de présenter les syntagmes de niveau quatre à partir du choix d'un syntagme de troisième niveau et de la demande par l'utilisateur de rechercher ceux de niveau quatre. Il utilise la requête Requête sur les SN 4. Le formulaire permet grâce à différents boutons d'option de :

- Chercher les syntagmes de cinquième niveau La procédure associée à ce bouton, cherche les syntagmes de niveau cinq et ouvre le formulaire **Voir syntagmes niveau cinq** ;
- Voir les titres une fois choisi un syntagme de niveau quatre La procédure ouvre le formulaire **Voir les articles 4** ;
- Quitter le formulaire courant vers celui appelé **Voir syntagmes niveau trois** Si cette option est choisie, l'application ferme le formulaire courant ;
- Faire une nouvelle recherche à partir d'un mot ou d'un centre de syntagme Si cette option est choisie, l'application ferme le formulaire courant, le formulaire **Voir syntagmes niveau trois** et le formulaire **Voir syntagmes niveau deux** ;
- Quitter l'application Si cette option est choisie, l'application ferme le formulaire courant, le formulaire **Voir syntagmes niveau trois**, le formulaire **Voir syntagme niveau deux** et le formulaire **Formreq**.

Voir syntagmes niveau cinq Ce formulaire est constitué d'un sous-formulaire, 1. syntagmes de cinquième niveau, qui a la fonction de présenter les syntagmes de niveau cinq à partir du choix d'un syntagme de quatrième niveau et de la demande par l'utilisateur de rechercher ceux de niveau cinq. Il utilise la requête Requête sur les SN 5. Le formulaire a comme boutons d'option :

- Voir les titres une fois choisi un syntagme de niveau cinq ; La procédure ouvre le formulaire **Voir les articles 5** ;
- Quitter le formulaire courant vers le formulaire **Voir syntagmes niveau quatre** Si cette option est choisie, l'application ferme le formulaire courant ;
- Faire une nouvelle recherche à partir d'un mot ou d'un centre de syntagme Si cette option est choisie, l'application ferme le formulaire courant, le formulaire **Voir syntagmes niveau quatre**, le formulaire **Voir syntagmes niveau trois** et le formulaire **Voir syntagmes niveau deux** ;

- Quitter l'application Si cette option est choisie, l'application ferme le formulaire courant, le formulaire **Voir syntagmes niveau quatre,** le formulaire **Voir syntagmes niveau trois**, le formulaire **Voir syntagmes niveau deux** et le formulaire **Formreq**.

Voir les articles 1, Voir les articles 2, Voir les articles 3, Voir les articles 4 et Voir les articles 5 Ces formulaires utilisent des sous-formulaires pour montrer les titres des articles ; ces sous-formulaires sont respectivement : Montre Articles 1, Montre Articles 2, Montre Articles 3, Montre Articles 4 etMontre Articles 5. Les sous-formulaires ont la fonction de rechercher les titres des articles et de les présenter dans le formulaire correspondant. Les requêtes responsables pour ces recherches sont respectivement pour chacun des sous-formulaire : Requête pour voir les titres 1, Requête pour voir les titres 2, Requête pour voir les titres 3, Requête pour voir les titres 4 et Requête pour voir les titres 5. Ces formulaires ont comme boutons d'option : 1.

- Voir l'article selon le choix du titre par l'utilisateur La procédure ouvre le formulaire Montre Doc ;
- Quitter le formulaire vers le dernier formulaire ouvert L'application exécute la procédure de fermeture du formulaire courant.

Montre Doc La source pour ce formulaire est la table Articles. Ce formulaire a comme objectif de montrer l'article selon le choix du titre par l'utilisateur. 1.

Formulaire Saisir Syntagmes Les objectifs de ce formulaire sont : réviser les syntagmes extraits du corpus, construire l'arborescence des syntagmes nominaux et inclure des nouveaux syntagmes nominaux. La source pour ce formulaire est la Table Gros Index. L'utilisation de ce formulaire a été très importante pour le chargement de la base de données et la construction de l'arborescence des syntagmes nominaux, parce qu'il a permis de faire : a) la révision des syntagmes nominaux extraits ; b) l'association entre les syntagmes ; c) l'attribution de centres de syntagmes à chaque syntagme nominal de premier niveau ; enfin d) l'attribution de niveaux aux syntagmes en observant le caractère relatif d'association entre deux syntagmes donnés. 2.

7. Conclusion

Les procédures de développement et de mise en service de la maquette se sont effectuées suivant deux étapes : la première a été une étape d'expérimentation où une petite maquette a été construite suivant un modèle de données similaire au modèle final présenté dans ce chapitre. Pendant cette phase : la mise en relation entre les syntagmes nominaux a été faite directement entre les suites de caractères des syntagmes nominaux et non par l'intermédiaire de codes d'identification ; la mise en arbre des syntagmes nominaux a considéré les niveaux absolus des syntagmes et les syntagmes de premier niveau ont été mis en relation directe avec les centres des syntagmes. On a chargé cinq articles dans la base de données pour tester cette maquette. Cette étape était importante pour connaître les limitations du logiciel Access et du modèle d'arborescence initial. Les

problèmes soulevés dans cette phase et les solutions adoptées pour construire la maquette définitive vont être présentés dans le chapitre suivant.

Dans la deuxième étape on s'est occupé de la construction de la maquette définitive en tenant compte les restrictions du logiciel et les solutions trouvées à propos de l'arborescence des syntagmes nominaux. Le développement présenté dans ce chapitre concerne la maquette définitive.

Cette démarche a été très utile dans le processus d'apprentissage du logiciel Access et étant donné la limitation de temps, on a permis d'apprendre en l'utilisant.

« Une expérience scientifique est [...] une expérience qui contredit l'expérience commune. » Bachelard (Gaston), La Formation de l'esprit scientifique (Vrin).

Chapitre 5 Mise en service de la maquette

1 Considérations préliminaires

Nous avons construit deux versions de la maquette. La première a permis de mieux connaître la mise en place d'un arbre de syntagmes nominaux. A partir des remarques obtenues lors de la mise en opération de cette première maquette, nous avons élaboré une deuxième version en l'améliorant grâce à celles-ci. Nous allons présenter ici les remarques relevées en exploitant la première maquette ; ce qui a permis de connaître aussi les limitâtes du logiciel Access.

2 Chargement de la base de données dans la maquette

Pour la première étape de la construction et de la mise en service de la maquette du système de recherche d'information, nous n'avons chargé que 5 articles dans la base de données.

La procédure de chargement des cinq premiers articles dans la base de données a été exécutée comme suit :

1. Identification des niveaux des syntagmes nominaux dans les fichiers Word 6.0a ;
2. Importation des syntagmes nominaux, en format Word 6.0a dans la table de travail GROSIND ;
3. Création de la table GROS INDEX à partir de la table GROSIND ;
4. Détermination des centres des syntagmes nominaux pour chaque syntagme de premier niveau et création de la table de Centre du Syntagme ;
5. Création des tables des syntagmes, tables des syntagmes niveau 1, 2, 3, 4 et 5 à partir de requêtes de sélection sur la table Gros Index ;

6. Construction de l'arborescence des syntagmes nominaux à l'aide de la création des tables de liaison entre les syntagmes nominaux d'un niveau donné avec son correspondant de niveau inférieur (syntagmes nominaux niveau 2 avec syntagmes nominaux niveau 1, syntagmes nominaux niveau 3 avec les correspondants niveau 2 et ainsi de suite). Pour établir cette liaison on a créé un formulaire pour chaque association ;
7. Création des tables de références et des articles ;
8. Comptage du nombre d'articles d'où chaque syntagme a été extrait. Ce comptage a été fait en utilisant des requêtes de sélection et des requêtes d'ajout.

Le travail pour inclure les cinq premiers articles dans la base de données et de construire l'arborescence des syntagmes nominaux a été très lourd, étant donné que toute la procédure était manuelle et que la construction de chaque niveau d'arborescence prenait en compte un seul syntagme à la fois et ceci à chaque niveau.

A partir de cette expérience, pour le chargement définitif du corpus dans la base de données, on a adopté les procédures suivantes :

1. Importation des fichiers Word 6.0a, contenant les syntagmes nominaux, groupés par chaque article, dans la table de GROSIND ;
2. Création, à partir de la table GROSIND, de la table GROS INDEX, mettant le champ des syntagmes nominaux, tant dans le champ syntagme que dans le champ syntagme nominal inférieur ; cette procédure a évité la tâche de saisir à nouveau manuellement chaque syntagme nominal de niveau inférieur
3. Vérification, à l'aide du formulaire Saisir Syntagmes, de tous les syntagmes nominaux. Grâce à cette procédure nous avons corrigé le champ « syntagmes nominaux inférieur », étant donné que ce champ a été créé à l'image du champ « syntagme ». Avec cette révision, nous avons défini aussi le niveau relatif d'association entre les syntagmes nominaux. Le développement de ce formulaire a permis de rendre la tâche de construction de l'arborescence et de définition des centres des syntagmes nominaux moins lourde que dans l'expérimentation initiale
4. Introduction des flexions en nombre des centres de syntagme nominal au moyen du formulaire « X table centre du syntagme », qui est à l'origine de la table des mots ;
5. Création de toutes les tables définies dans la maquette, au moyen des requêtes de sélection et d'ajout, à partir de la table Gros Index.

L'expérimentation de la maquette avec les cinq premiers articles a permis de se rendre compte des limites suivantes au sujet du logiciel : a) la taille maximale d'un champ type texte est de 256 caractères ; b) le logiciel n'arrive pas à travailler correctement avec une requête d'ajout dont la somme des tailles des champs soit est supérieur à 256 caractères ; c) la recherche d'un champ type texte est plus lente que n'importe quel autre type de champ. Parmi ces limitations, la plus importante est celle du nombre de caractères (256). Ceci empêche la liaison de deux champs ou plus, alors que ce type d'opération est très commun dans une procédure de recherche d'information. Pour éviter ces problèmes, dans la maquette finale, nous avons créé un code unique pour chaque syntagme nominal.

Ainsi toutes les opérations de comparaison et d'ajout sont effectuées sur le code et non pas sur le texte du syntagme nominal. Ainsi pour restreindre la longueur d'un champ nous avons décidé de limiter sa taille à 150 caractères.

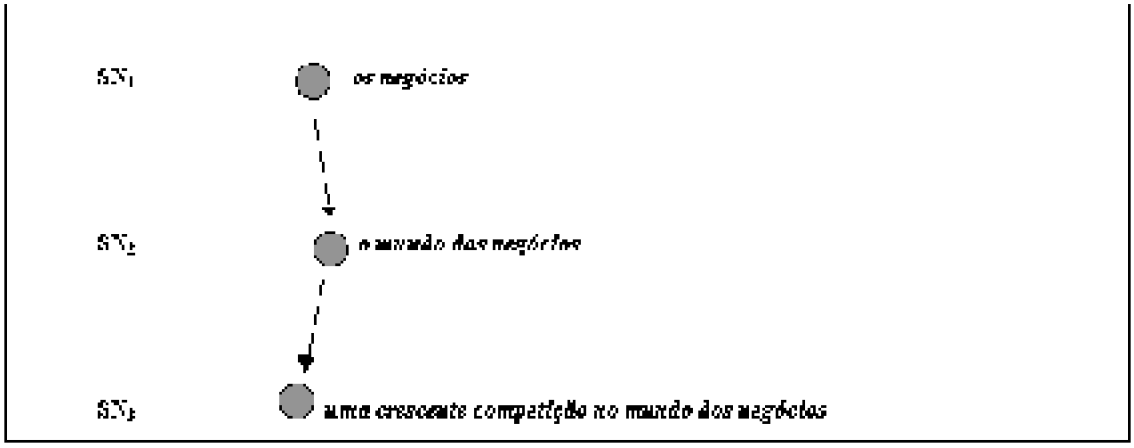
Cette limitation n'a offert que deux solutions pour stocker les textes des articles. Une solution étant de les considérer comme un objet importé, une autre de les mettre dans un champ type mémo. Aucune de ces deux solutions n'était la bonne, car elles ne permettaient pas de traiter les textes. Pour la maquette il fallait avoir des possibilités de distinction des syntagmes nominaux dans les textes lorsqu'on demande de voir le contenu d'un article. Ainsi, parmi les deux solutions la seconde étant la moins contraignante, on a gardé donc les textes des articles dans les champs type mémo. Cela a permis de présenter l'article en entier, ce qui avait été impossible autrement.

Les problèmes relatifs au comportement des syntagmes nominaux dans leur organisation en arbre et aux centres des syntagmes seront discutés dans la section suivante.

3 Comportement des syntagmes nominaux dans l'organisation en arbre

L'approche initiale pour la construction de l'arborescence a été établie en considérant l'organisation des syntagmes nominaux structurés en niveaux attribués de façon absolue les uns par rapport aux autres. Ainsi, on a créé des tables qui associent un syntagme nominal d'un niveau donné avec le syntagme nominal d'un rang immédiatement inférieur.

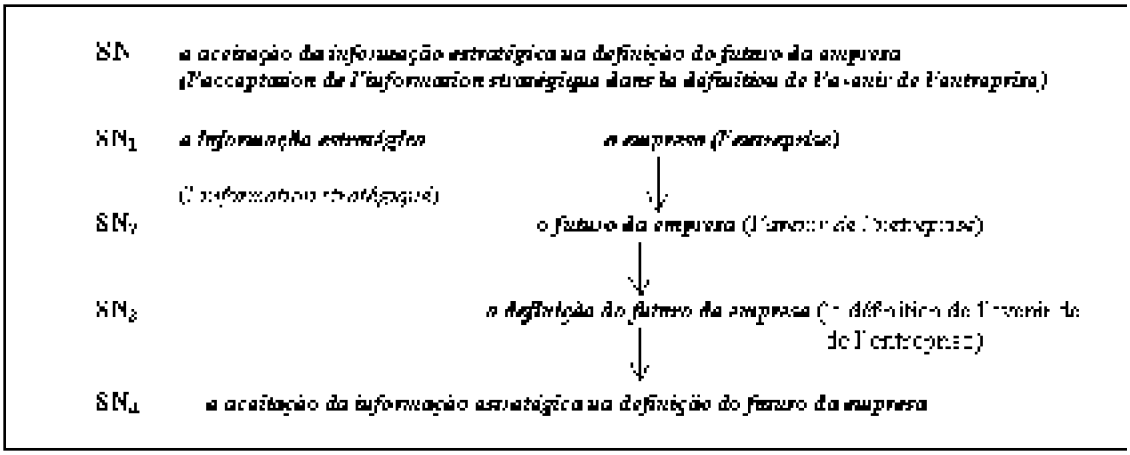
Par exemple, dans la table LIAISON SN 1 - SN 2 nous avons associé à chaque syntagme nominal de deuxième niveau tous les syntagmes de premier niveau d'où ils ont été extraits. Pour la table LIAISON SN 2 - SN 3, on a associé à chaque syntagme de troisième niveau tous les syntagmes de deuxième niveau d'où ils ont été extraits et ainsi de suite. D'une manière générale, on peut représenter graphiquement cette arborescence comme dans la figure 5.1.



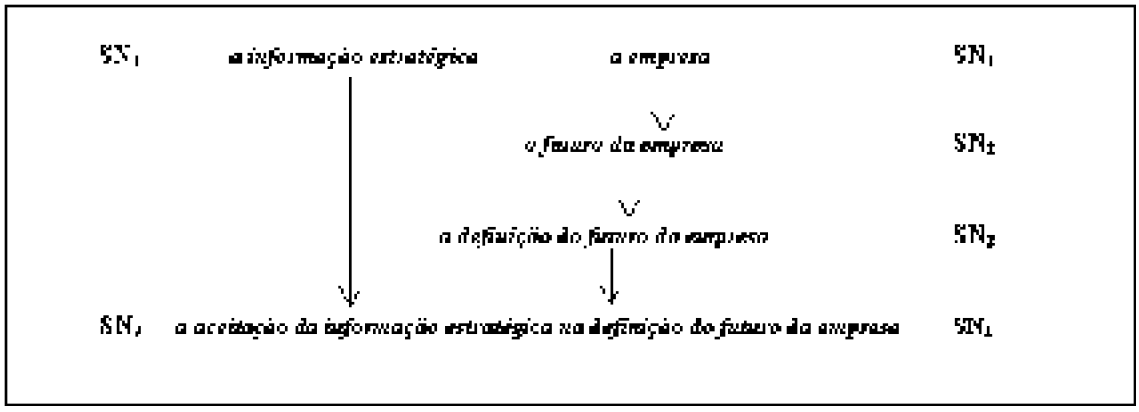
Toutefois, certains syntagmes nominaux n'ont pas la possibilité d'être associés au syntagme nominal d'où ils ont été extraits. Ce sont les cas des syntagmes nominaux avec double rection. Chaque rection donne naissance à une branche indépendante de

syntagmes nominaux. Si des deux branches proviennent de syntagmes nominaux de niveaux différents, la question est : quel niveau doit-on attribuer au syntagme d'où ils sont extraits ? Au départ on a adopté la solution d'attribuer le niveau absolu. C'est-à-dire, lors de l'extraction, l'attribution de niveau à chaque syntagme nominal a été faite en tenant en compte de l'enchaînement le plus long de syntagmes nominaux.

Dans l'exemple de la figure 5.2 on montre ce problème. Cette approche n'a pas fonctionné car on n'arrive au syntagme nominal de plus haut niveau que par le centre de syntagme nominal de la branche dont l'enchaînement des syntagmes nominaux est le plus long. Dans la figure 5.2 on peut mieux voir ce problème. Dans cet exemple on n'a pas pu arriver au syntagme de plus haut niveau au moyen du syntagme de premier niveau ***a informação estratégica*** *(l'information stratégique)*.



Pour résoudre ce problème, au lieu d'organiser l'arborescence en considérant les niveaux de chaque syntagme nominal de manière absolue, nous avons construit l'arborescence en gardant les niveaux relatifs d'association entre les syntagmes nominaux.



Dans la figure 5.3, on organise l'arborescence en considérant les niveaux relatifs d'associations entre les syntagmes nominaux. Le syntagme nominal *a aceitação da informação estratégica na definição do futuro da empresa* est, dans cette nouvelle approche, à la fois syntagme nominal de deuxième niveau par rapport au syntagme nominal *a informação estratégica* et syntagme nominal de quatrième niveau par rapport au syntagme nominal *a empresa*.

Ce changement n'altère en rien la maquette du système de recherche d'information, étant donné que la construction de l'arborescence des syntagmes nominaux est faite manuellement.

4 Centres complémentaires des syntagmes nominaux

Lorsqu'une recherche d'information était faite dans la première maquette, cherchant toujours l'information à partir des centres des syntagmes nominaux. Or, nous avons constaté qu'ils ne sont pas suffisant pour trouver l'information, car on risque d'avoir quelque taux de silence (nombre de références ou documents pertinents manqués à la suite d'une recherche d'information, alors qu'ils existent dans la base de données). Usuellement cela arrive avec les syntagmes nominaux composés d'une expansion prépositionnelle, comme par exemple : *Os sistemas de informação.*

Le centre du syntagme nominal est : *sistemas.* Or, bien que le mot *informação*, dans ce cas, ne soit pas le centre du syntagme, il est quand même important dans la recherche d'information. Lorsqu'on fait la recherche à partir du centre du syntagme nominal *informação,* on ne trouve pas les documents indexés par le syntagme nominal *os sistemas de informação*. Cela produit des taux de silence. Pour résoudre ce problème on propose la création d'une figure de « centre complémentaire des syntagmes nominaux ». Ce sont des mots qui ont une importance égale aux centres des syntagmes nominaux.

Du point de vue linguistique ces types de syntagmes nominaux (les systèmes d'information, le stockage d'information) sont réécrit comme étant.

1. N + P + N Exemple : système d'information Où : N est un nom, un prédicat libre. C'est-à-dire, autant le mot système que le mot information sont des prédicats libres. P est une préposition.
2. EP = P + N EP est une expansion prépositionnelle Ainsi, on peut réduire (1) à :
3. N + EP Et finalement N + EP est égal à N (selon la grammaire développé par le groupe SYDO)
4. N □ N + EP

Etant donné qu'il s'agit de trois N, tous des prédicats libres et que ce mot est aussi un mot composé, il nous semble raisonnable de prendre aussi les deux autres N comme une sorte de centres de syntagme nominal (dans le cas *Systèmes d'information*, le centre du syntagme est *systèmes* et les mots *information* et *systèmes d'information* jouent aussi le rôle de centre du syntagme nominal). Ce qu'on nomme centres complémentaires de syntagme nominal de premier niveau.

Du point de vue de la maquette, il faut créer une structure capable de permettre la recherche non seulement à partir des centres de syntagmes nominaux, mais aussi à partir des centres complémentaires des syntagmes nominaux. Pour cela, il y a deux solutions possibles : a) créer une table de mots complémentaires composés de mots qui ne sont pas des centres de syntagmes nominaux, mais qui sont quand même très importants pour la recherche ; b) inclure ces mots dans la TABLE CENTRE DU SYNTAGME.

La solution 'a' est plus intéressant du fait que la TABLE CENTRE DU SYNTAGME resterait intègre. Or, ce type de solution est cependant le moins performant car le système doit faire la recherche dans deux tables au lieu de la faire dans une seule.

La solution 'b' qui est moins intéressante du point de vue de la structure de données, montre qu'on pourra avoir des mots dans la TABLE CENTRE DU SYNTAGME qui ne sont pas vraiment des centres des syntagmes nominaux. Par contre, du point de vue de la performance du système de recherche d'information, c'est la solution la plus indiquée, car le système ne fera alors la recherche que dans une seule table.

5 Centres des syntagmes nominaux et ses flexions

Dans la première version du système de recherche d'information on a laissé les centres des syntagmes nominaux tels qu'ils sont apparus dans les syntagmes et dans le corpus. On s'est aperçu, en exploitant le système, que les résultats étaient faibles lorsqu'on essayait de chercher des syntagmes nominaux dont le centre était *informação*. Le système trouvait *a informação, a informação científica, a informação técnica, etc.* Or, le système ne trouvait pas des syntagmes comme : *as informações, as informações científicas, as informações industriais, as informações organizacionais, as informações técnicas, etc.* On avait encore de taux de silence !

Pour résoudre ce problème nous avions considéré deux solutions possibles : 1) la mise en œuvre des opérateurs de troncature ; et 2) le traitement des flexions, en nombre et en genre, des centres de syntagmes nominaux en créant une table de mots qui soit équivalente aux centres des syntagmes nominaux.

La solution de mise en œuvre des opérateurs de troncature n'est pas bonne car on il existe des mots dont la flexion n'est pas faite à la fin. Cette solution ne peut donc résoudre que partialement le problème. C'est pourquoi nous avons choisi la deuxième solution.

Quoi qu'il en soit, dans un système d'information, nous trouvons important de conserver les deux solutions, puisqu'elles rendront le système plus souple et les résultats auront moins de silence ; dans cette étape de la recherche d'information, l'augmentation du « bruit » (proportion de références ou documents non-pertinents trouvés à l'aide d'une recherche d'information) est plus intéressante et souhaitable qu'un taux élevé de silence. Cela se justifie parce que l'usager, à ce moment de la recherche, est en train de choisir les syntagmes qui appartiennent encore au premier niveau et cela suppose un plus grand choix. On voit ici l'importance de l'interaction entre le système de recherche d'information et l'utilisateur.

Après l'implémentation de la solution choisie et l'exploitation de la maquette en vérifiant la navigation dans l'arborescence des syntagmes nominaux, nous avons observé l'apparition de plusieurs syntagmes nominaux de même signification mais ayant une composition différente ou une flexion en nombre. Exemple : à partir d'une demande utilisant le mot *informação* en tant que centre de syntagme nominal, on trouve des syntagmes nominaux comme *a informação, as informações, informações* parmi d'autres syntagmes nominaux associés à ce centre de syntagme nominal. Si l'on veut connaître

tous les syntagmes nominaux de deuxième niveau associés au syntagme nominal *a informação*, il faut chercher aussi les syntagmes nominaux associés à *as informações* et *informações* aussi, étant donné qu'ils font référence au même sujet. Cependant la maquette n'offrait pas cette possibilité.

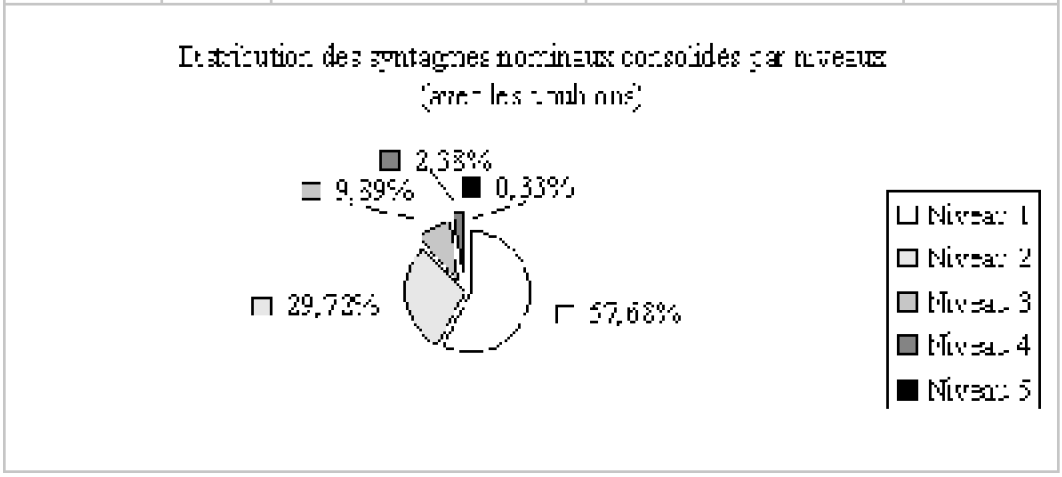
Une solution possible, serait de développer une simulation de l'opérateur booléen « ou », étant donné que l'interface développé est orientée par menus.

6 Statistique descriptive des syntagmes nominaux

Les associations et comportements des syntagmes nominaux dans l'arborescence peuvent être vus au moyen des statistiques descriptives. La figure 5.4, montre la distribution des syntagmes nominaux dans les quinze (15) articles avec toutes ses occurrences.

Il faut dire que l'occurrence multiple d'un même syntagme nominal résulte non seulement de l'occurrence naturelle dans les articles, mais du calcul des anaphores et des syntagmes nominaux avec factorisation. Ces calculs ont produit plusieurs syntagmes nominaux. C'est pour cela qu'on trouve parfois la répétition des syntagmes nominaux dans un même paragraphe.

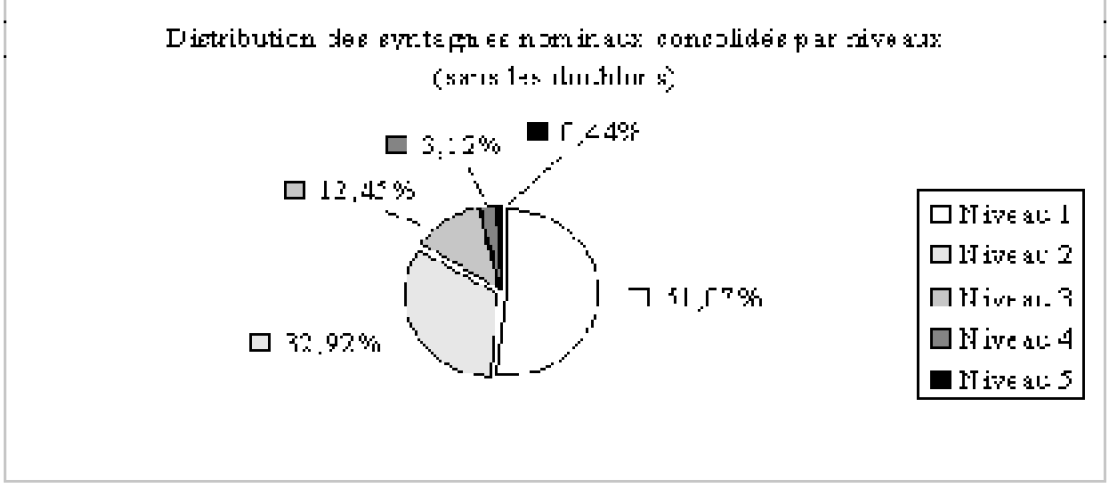
Article	Niveau 1	Niveau 2	Niveau 3	Niveau 4	Niveau 5	Total
1	378	221	63	22	7	681
2	357	161	52	6	1	577
3	338	152	50	19		559
4	290	153	47	7	3	502
5	320	136	47	23		726
6	348	78	87	9		588
7	368	175	61	20		624
8	471	211	91	24	9	806
9	288	200	76	21	5	590
10	266	150	50	12		478
11	267	119	41	10		467
12	362	194	46	9		611
13	278	168	66	6	1	519
14	396	209	53	13	1	672
15	352	165	54	13	2	586
Total	5086	2621	872	210	29	8818



La figure 5.5, montre la distribution des syntagmes nominaux dans le corpus, sans les doublons. Il faut dire que, bien que dans ce tableau les doublons des syntagmes nominaux dans un même article n'apparaissent pas, on peut trouver encore des doublons des syntagmes nominaux parmi l'ensemble des syntagmes nominaux de plus d'un article.

La comparaison de ces deux tableaux (figures 5.4 et 5.5), indique une chute, en pourcentage, des syntagmes nominaux de niveau 1 et une augmentation, aussi en pourcentage, des syntagmes nominaux des autres niveaux. Cette augmentation s'explique par une quantité plus grande de doublons des syntagmes nominaux de niveau 1 par rapport à ceux des autres niveaux. On voit que les doublons des syntagmes nominaux de niveaux plus élevés (3, 4 et 5) sont plus rares que ceux de niveaux moins élevés (1 et 2).

Article	Niveau 1	Niveau 2	Niveau 3	Niveau 4	Niveau 5	Total
1	233	184	49	37	7	507
2	284	142	51	6	1	484
3	177	125	45	15		362
4	154	131	47	9	3	344
5	396	134	47	23		590
6	170	139	86	8		403
7	259	139	57	20		475
8	197	173	79	22	7	484
9	197	179	71	19	5	471
10	179	113	48	13		353
11	165	119	40	10		334
12	274	176	43	9		502
13	198	139	62	6	1	406
14	266	168	51	13	1	499
15	233	130	50	13	2	428
Total	3389	2184	826	207	29	5635



Le pourcentage de doublons des syntagmes nominaux, dans chaque niveau a été : a) niveau 1, 50,07% ; b) niveau 2, de 20,01% ; c) niveau 3, de 5,57% ; d) niveau 4, de 1,45% ; et e) niveau 5, de 0%.

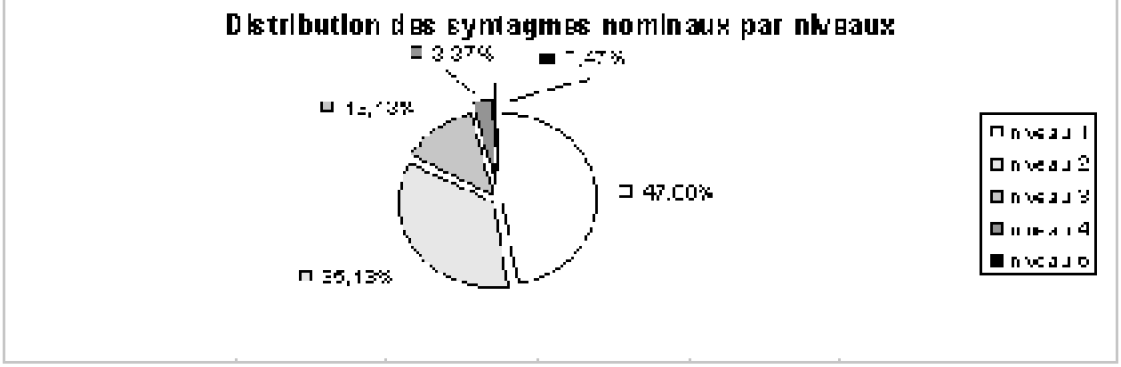
Depuis la construction de la base de données, où nous avons créé des tables pour chaque niveau des syntagmes nominaux, il a été alors possible de connaître leur distribution finale sans doublons, c'est-à-dire le nombre des syntagmes nominaux uniques pour chaque niveau. La figure 5.6 illustre la distribution des syntagmes nominaux par niveau, sans aucun doublon.

Comme dans la comparaison entre le tableau Distribution des syntagmes nominaux avec doublons et le tableau Distribution des syntagmes nominaux sans doublons dans chaque article, la figure 5.6 indique un petit accroissement dans le pourcentage des syntagmes nominaux des niveaux 2, 3, 4 et 5, en opposition à la chute du pourcentage des syntagmes nominaux de premier niveau. Dans les deux cas, la quantité de doublons des syntagmes nominaux de premier niveau est plus grande que celle des autres

niveaux. D'ailleurs, il n'y a pas eu de doublons sur les syntagmes nominaux des niveaux 3, 4 et 5.

La constatation de l'inexistence des doublons des syntagmes nominaux, à partir des niveaux 3, 4 et 5, entre les articles du corpus est cohérente avec l'idée selon laquelle ces niveaux sont responsables du raffinement de la recherche d'information.

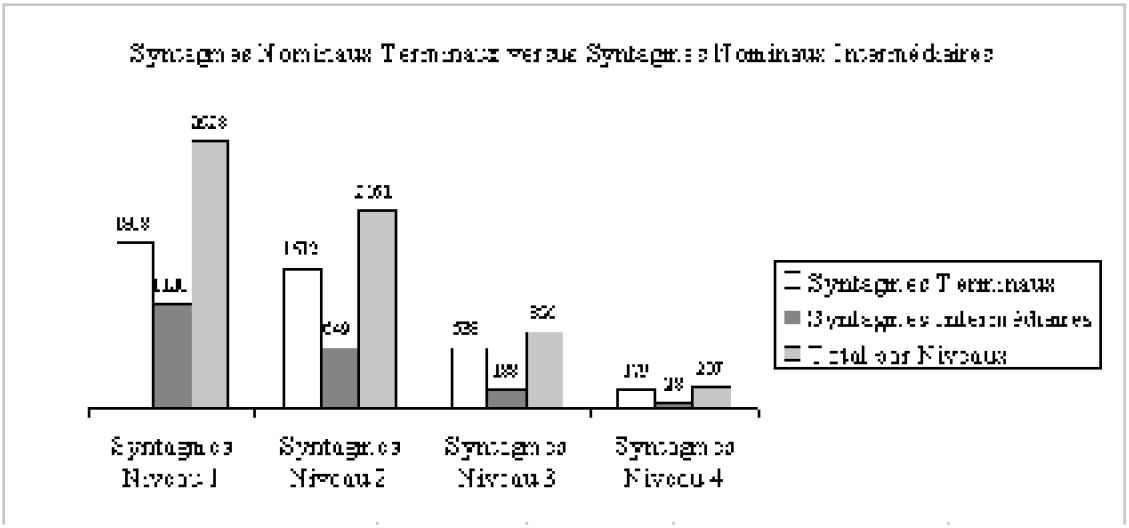
	Syntagme niveau 1	Syntagme niveau 2	Syntagme niveau 3	Syntagme niveau 4	Syntagme niveau 5	Total
N°. de syntagmes	2923	2161	826	207	29	6151



Dans l'arborescence on trouve deux genres de syntagmes nominaux ; les premiers qui ne sont pas associés à aucun syntagme nominal, et qu'on appellera désormais syntagmes nominaux terminaux, et les deuxièmes qui se trouvent associés aux syntagmes nominaux de niveau supérieur, que nous appellerons syntagmes intermédiaires.

	NIVEAU 1		NIVEAU 2		NIVEAU 3		NIVEAU 4		TOTAL	
	N°Syn.	%	N°Syn.	%	N°Syn.	%	N°Syn.	%	N°Syn	%
Syn. Terminaux	1803	61,70%	1523	70,47%	638	77,24%	177	85,51%	4141	67,32%
Syn. Intermédiaires	1120	38,30%	638	29,53%	188	22,76%	30	14,49%	1976	32,12%
Total	2923	100,00%	2161	100,00%	826	100,00%	207	100,00%	6117	100,00%

Les figures 5.7 et 5.8 montrent la quantité des syntagmes nominaux terminaux et celle des intermédiaires.

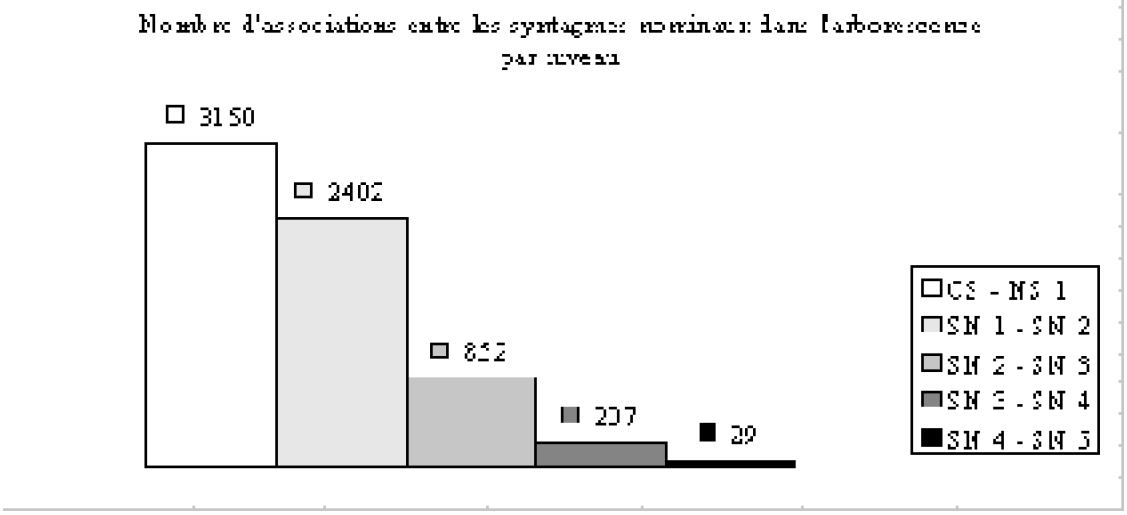


Ainsi, on a 1808 syntagmes nominaux de premier niveau qui sont aussi terminaux parce qu'ils ne sont associés à aucun syntagme nominal de niveau deux. Comme syntagmes nominaux intermédiaires de premier niveau on a 1120 syntagmes nominaux associés aux syntagmes nominaux de deuxième niveau. Par rapport aux syntagmes nominaux de niveau deux, 1512 sont terminaux et 649 sont associés aux syntagmes nominaux de troisième niveau. La même interprétation est donnée aux syntagmes nominaux de niveau trois et quatre. On se rend compte qu'il y a une décroissance d'environ 8% sur le nombre des syntagmes nominaux d'un niveau donné par rapport à ceux d'un niveau immédiatement supérieur. Ce fait démontre la capacité de raffinement que l'arborescence des syntagmes nominaux possède dans la procédure de navigation et de recherche d'information. Cependant, il faut faire une remarque à l'égard des syntagmes nominaux terminaux et intermédiaires : même les intermédiaires pourront être des syntagmes nominaux terminaux dans certains documents, tandis que dans d'autres ils pourront être seulement intermédiaires. On explique cela par le fait que dans une base de données, on peut trouver des documents qui parlent d'un sujet de manière plus spécifique et d'autres qui en parlent de manière plus générale.

Par rapport à l'arborescence des syntagmes nominaux on a construit le tableau de la figure 5.9.

La figure 5.9 montre qu'il y a 3150 associations entre les centres de syntagmes nominaux et les syntagmes nominaux de premier niveau, 2402 associations entre les syntagmes nominaux intermédiaires de premier niveau et les syntagmes nominaux de deuxième niveau, 852 associations entre les syntagmes nominaux intermédiaires de deuxième niveau et les syntagmes nominaux de troisième niveau, et ainsi de suite.

	CS - NS 1	SN 1 - SN 2	SN 2 - SN 3	SN 3 - SN 4	SN 4 - SN 5	Total
N° Associations	3150	2402	852	207	29	6640
N° Relatif d'As	47,44%	36,17%	12,83%	3,12%	0,44%	100,00%



La figure 5.9 montre qu'il y a 3150 associations entre les centres de syntagmes nominaux et les syntagmes nominaux de premier niveau, 2402 associations entre les syntagmes nominaux intermédiaires de premier niveau et les syntagmes nominaux de deuxième niveau, 852 associations entre les syntagmes nominaux intermédiaires de deuxième niveau et les syntagmes nominaux de troisième niveau, et ainsi de suite.

L'analyse de ce tableau montre, encore une autre fois, la capacité de raffinement permis par l'arborescence des syntagmes nominaux dans une procédure de navigation et de recherche d'information. On constate que dans les associations entre les centres de syntagme nominal et les syntagmes nominaux de premier niveau aussi bien qu'entre les syntagmes nominaux intermédiaires de premier niveau et les syntagmes nominaux de deuxième niveau et entre les syntagmes nominaux intermédiaires du deuxième niveau et ceux de troisième niveau, un syntagme nominal donné amène à plusieurs syntagmes nominaux de niveau supérieur et vice-versa. Ce fait justifie la création, dans la structure de données de la base, des tables spécifiques pour les associations dont les clés composantes sont le code des syntagmes nominaux de niveau inférieur et le code des syntagmes nominaux de niveau supérieur. Cela évite les doublons de clés dans une table.

7 Conclusion

Bien que des statistiques descriptives sur les syntagmes nominaux et leurs comportements aient permis des constatations précises, il est nécessaire de procéder à une évaluation de la maquette et du comportement des syntagmes nominaux en tant que structure d'accès à l'information, avec la participation des utilisateurs. Or, les délais imposés pour la mise en place de ce travail n'a pas rendu possible l'exécution de cette tâche. Cependant, la maquette est prête pour être soumise à l'évaluation proposée. Étant donné la façon dont la maquette a été développée, la construction d'une autre base de

donnés quelconque, même dans une autre langue que la langue portugaise est tout à fait possible. Sachant qu'il faut construire une base de données important puisqu'une évaluation demande un volume plus grand que celui qui nous avons traité dans cette recherche. Enfin, la maquette peut être un outil d'expérimentation très important pour l'évaluation de cette approche, mais il faut avoir une procédure automatisée de reconnaissance, extraction et indexation de syntagmes nominaux.

« L'observation est l'investigation d'un phénomène naturel, et l'expérience est l'investigation d'un phénomène modifié par l'investigateur. » Bernard (Claude) , Introduction à l'étude de la médecine expérimentale.

Chapitre 6 Exploitation de la maquette

1 Considérations préliminaires

L'utilisation des syntagmes nominaux comme moyen d'accès à l'information dans une base de données textuelles est-elle efficace ? C'est, en effet, une bonne et importante question à être répondue, étant donné que nous avons construit une maquette en utilisant cette approche. Nous sommes d'accord qu'il faut évaluer cette approche auprès des usagers. Cependant, ce n'est pas le moment de le faire. Il faut d'abord connaître les comportements des syntagmes nominaux dans une procédure de recherche d'information ainsi que le comportement de la maquette en tant qu'interface de recherche d'information. Ces connaissances seront importantes dans la construction d'un vrai système de recherche d'information utilisant l'approche proposée dans ce travail. En effet, l'évaluation de l'efficacité des syntagmes nominaux en tant que moyen d'accès à l'information ne peut être réalisée que par un tel système. La maquette est vu donc plus comme un outil d'acquisition de connaissances sur l'approche proposée. Dans ce sens là, nous exploiterons la maquette que nous avons construit dans le cadre du DEA en vue d'acquérir des connaissances sur le comportement des syntagmes nominaux en tant que descripteur ainsi que le comportement de la maquette comme interface de recherche d'information.

Nous ayons utilisé le thesaurus Tesauro Ciência da Informação [57] (TCI) - Thesaurus Science de l'Information - pour essayer de faire des recherches d'information suivant les termes qui y sont. L'utilisation de ce thesaurus est une manière de simuler la façon dont un utilisateur commun, du domaine des sciences de l'information, fait une recherche d'information.

Les termes dans le TCI sont structurés selon 7 catégories :

[57] CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO (CNPq). INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA (IBICT). *Tesauro Ciência da Informação (Versão Preliminar).* Brasília, 1989.

1. Information : sont regroupés, dans cette catégorie, les termes qui font référence aux aspects historiques, théoriques, sociaux, légaux et philosophiques liés à l'information ;
2. Document : sont regroupés, dans cette catégorie, les aspects qui font référence au type d'enregistrement de l'information et les supports où elle est enregistrée ;
3. Unités d'information : on regroupe, dans cette catégorie, les termes qui font référence aux institutions spécialisées dans l'acquisition, le traitement, le stockage et la dissémination d'informations ;
4. Planification, organisation et administration d'unités d'information : sont regroupés, dans cette catégorie, les termes relatifs aux activités liées à la conception et au fonctionnement des unités d'information ;
5. Processus et services d'information : dans cette catégorie on trouve les termes qui font référence aux activités de stockage, traitement et recherche d'information et les services offerts à partir de ces activités ;
6. Echange et utilisation d'information : sont placés dans cette catégorie les termes qui font référence aux aspects liés à l'étude du phénomène de la communication de l'information dès sa genèse jusqu'à son utilisation ;
7. Profession : on regroupe dans cette catégorie les aspects liés à l'enseignement, à l'apprentissage, à la formation, à l'entraînement, à l'éthique et au rôle du professionnel de l'information.

On a utilisé la liste alphabétique de termes de chaque catégorie dans la recherche d'information pour l'exploitation de la maquette. Cependant, on n'a pas exploité la catégorie G car nous n'avons pas disposé de la liste alphabétique de cette catégorie.

2 Exploitation de la maquette à l'aide du TCI

L'exploitation de la maquette a consisté à faire une recherche à partir de chaque terme de chaque catégorie sans le souci de vérifier si le domaine d'une catégorie existait dans le corpus ou non.

Dans le TCI certains termes sont considérés comme étant des descripteurs et d'autres qui ne le sont pas. Nous avons effectué des recherches pour tous les termes indistinctement.

Comme critère de recherche nous avons adopté de demander une information à partir du centre de syntagme nominal de premier niveau. Les termes du TCI sont des descripteurs, ils ne sont pas précédés de déterminants comme les syntagmes nominaux. Ainsi, pour faire l'analyse, on n'a pas tenu compte des déterminants existants dans les syntagmes nominaux trouvés.

Par exemple : On a trouvé le syntagme nominal **outras obras de referência** et on l'a compté comme étant un résultat correct pour le terme **obra de referência**. Dans ce cas on a, en plus de la question du déterminant, le fait que le terme est au singulier. Pour une question de normalisation, dans le TCI tous les termes sont au singulier.

En ce qui concerne l'utilisation de la maquette on peut dire qu'elle a toujours offert une manière très simple de faire la recherche, sans les complications usuelles d'un système de recherche d'information orienté par un langage de commande. Pourtant, la maquette ne permettait pas d'arriver à un résultat d'une manière plus rapide car on est obligé de faire la recherche toujours à partir d'un centre de syntagme nominal de premier niveau et puis monter l'arborescence. Cela peut éventuellement ennuyer l'utilisateur spécialiste, surtout s'il connaît le terme exact de l'information dont il a besoin et si ce terme est un syntagme nominal de quatrième ou cinquième niveau. Il semble que l'approche de la maquette est plus orientée vers les usagers débutants parce qu'elle leur permet de connaître le sujet de la base et les termes qui y existent. Une solution à ce problème peut être la création d'une option permettant à l'utilisateur de faire la demande de la recherche d'information à partir d'un centre de syntagme nominal d'un niveau plus haut.

Catégorie	Nombre de Termes	Termes trouvés	Pourcentage de termes trouvés
A – Information	49	15	30,61 %
B – Document	154	15	9,74 %
C - Unités d'information (UI)	91	6	6,59 %
D – Planification, organisation et administration d'UI	110	20	18,18 %
E – Processus et services d'information	263	16	6,14 %
F – Echange et utilisation d'information	42	7	16,67 %
Total	771	83	10,76 %

Nous présentons le résultat de la recherche, dans le tableau de la figure 6.1. Bien qu'on a fait la recherche en utilisant les descripteurs et non-descripteurs, on n'a tenu compte que des descripteurs. Il faut dire que nous avons trouvé aussi quelques réponses pour les non-descripteurs.

Articles	A	B	C	D	E	F	G
La connaissance comme ressource stratégique des entreprises	*						
L'intelligence compétitive et la décision des entreprises	*					*	
L'économie de l'information	*			*	*		
L'information comme matière première stratégique	*			*			
L'information technique-économique : plus important que jamais	*			*		*	
Perspectives de l'agent de l'information dans le contexte brésilien							*
Les systèmes d'information : l'évolution de ces approches						*	
Consultation informatologique en réseau : une alternative pour ...					*		*
L'information pour l'industrie				*	*	*	
Interaction entre les entreprises ayant besoin d'information et la...				*	*		
L'utilisation de l'information dans l'industrie comme paradigme pour le ...				*		*	
L'information efficace dans l'entreprise					*	*	
La gestion de l'information : changement dans les profils professionnels						*	*
L'information : outil de domination et de soumission	*					*	
L'information : la clé pour la qualité totale				*	*		
Total	6	-	-	7	6	8	3

L'analyse des résultats obtenus dans l'exploitation de la maquette à l'aide du TCI doit tenir compte du fait que le corpus ne couvre pas toutes les catégories du TCI. D'autre part, en tenant compte de l'évolution croissante du domaine des sciences de l'information et du fait que la réalisation du TCI date de 1989, on peut penser que des syntagmes nominaux sont des descripteurs potentiels pour ce thesaurus.

Pour essayer de faire une analyse plus précise il faut classer les articles selon les catégories du TCI. Pour cela on a utilisé les informations du mémoire [58] sur les articles.

Dans la figure 6.2, on remarque que les articles ont été souvent inclus dans plus d'une catégorie. En conséquence, la somme totale d'occurrence d'articles dans les catégories est plus grande que la quantité des articles.

En faisant une comparaison entre les deux tableaux, on peut observer les faits suivants :

1. Le pourcentage des termes - des catégories A, D et F - trouvés dans la maquette est aussi fort que la quantité des articles classés dans chacune de ces catégories. On a trouvé 21,74% de descripteurs de la catégorie A tandis qu'on n'a que 6 articles classés dans cette catégorie. En ce qui concerne la catégorie D, on a trouvé 18,10% de descripteurs contre 7 articles, alors que pour la catégorie F on a trouvé 16,66% de descripteurs contre 8 articles. Cependant l'ordre de grandeur de la quantité d'articles dans chacune de ces catégories ne correspond pas à celles de ses pourcentages. On observe que dans la catégorie A, moins d'articles (6) ont utilisé un pourcentage de

[58] Hélio KURAMOTO. *Maquette d'un système de recherche d'information en utilisant des syntagmes nominaux.* Villeurbanne, 1995. Mémoire du DEA. École Nationale Supérieure des Sciences de l'Information et des Bibliothèques.

termes plus grand que les catégories D et F. Par contre, la quantité d'articles classés dans la catégorie F est plus grand que dans les deux autres. Pourtant, son pourcentage est plus faible que les deux autres. On observe encore que la quantité de termes trouvés pour cette catégorie (8) est beaucoup plus faible que celle obtenue pour la catégorie E (18). Pourtant, le pourcentage de termes trouvés pour la catégorie F est beaucoup plus fort que celui de la catégorie E. On peut donc supposer que la catégorie F est, dans le TCI, moins développée que les autres catégories (48 termes contre 69 pour la catégorie A, 116 pour la catégorie D). En ce qui concerne les catégories A et D, on se rend compte que la catégorie D possède une quantité important de termes (116 contre 69 termes de la catégorie A). Ce qui donne une valeur relative plus grande pour la catégorie A ;

2. Les catégories B et C ne tiennent aucun article classé comme tel, et pourtant on a trouvé quelques-uns uns de ses termes dans la maquette (9,74% et 6,59% respectivement). On a trouvé 15 termes du TCI dans la maquette pour la catégorie B (Document). Lorsqu'on parle d'une unité d'information ou de sa planification, son organisation ou son administration, on est obligé de parler aussi des périodiques, des articles, des brevets, des normes, etc. Ce sont des termes qui font partie de la catégorie B dans le TCI. On peut faire le même raisonnement pour la catégorie C qui regroupe des termes concernant les unités d'information comme base de données, bibliothèque, système d'information, etc ;

3. On trouve un nombre important d'articles classés dans la catégorie E (6) et pourtant la quantité relative de termes de cette catégorie, trouvés dans la maquette, n'est pas signifiante. C'est vraisemblablement à cause de la quantité totale importante de termes existants dans cette catégorie dans le TCI, 293. Cette catégorie possède presque 150% de plus de termes que la catégorie D (116 termes) tandis qu'on a trouvé 21 termes de cette catégorie et 18 de la catégorie E. On est amené à supposer que la catégorie E est plus développée que les autres.

D'une autre part nous observons quelques syntagmes nominaux qui possèdent une forme syntaxique un peu différente des termes du TCI. Ces syntagmes n'ont pas été comptés dans la statistique. Voilà quelques exemples :

- **demanda dos usuários** (dans la maquette) au lieu de **demanda de usuários** (dans le TCI) Bien que les deux termes se ressemblent, ils ont une sémantique différente. Le terme **demanda de usuários** fait référence à la demande des utilisateurs d'une manière générale tandis que le terme **demanda dos usuários** fait référence à un ensemble d'utilisateurs définis. On a dans le terme **demanda de usuários** l'intervention de la logique intensionnelle car on n'a pas forcement des utilisateurs. Par contre, dans le terme **demanda dos usuários** on a l'intervention de la logique extensionelle. C'est la mise en relation de la demande et d'un ensemble d'utilisateurs définis. Dans le terme **demanda dos usuários**, on a deux syntagmes nominaux (**demanda dos usuários** et **os usuários**), alors que dans le syntagme **demanda de usuários**, on n'a que un seul syntagme nominal ;
- **gerador de conhecimento** (dans la maquette) au lieu de **gerador de informação**

(dans le TCI) Ces deux termes sont peut-être des termes associés mais ils n'ont pas le même sens. On peut dire qu'un générateur de connaissance génère des informations mais le contraire n'est toujours pas vrai ;

- **estudo de mercado** (dans la maquette)au lieu de **estudo de demanda** (dans le TCI) On voit ici encore deux termes qui se ressemblent ou, au moins font partie d'un même domaine. Le terme **estudo de mercado** (étude de marché) est vraisemblablement plus général que **estudo de demanda** (étude de demande) ;
- **informatização na sociedade** (dans la maquette) au lieu de **informatização da sociedade** (dans le TCI) Ces termes font référence au même sujet, mais utilisent une syntaxe différente. Le TCI a établi le terme **informatização da sociedade** (informatisation de la société) pour faire référence à l'informatisation des activités existant dans la société. Par contre, il y a quelques auteurs qui appellent ce processus de **informatização na sociedade** (informatisation dans la société). Ce genre de problème montre en fait qu'on peut trouver de petites différences entre les termes établis dans un thesaurus et les termes utilisés dans les articles par les auteurs ; Dans un autre côté, chacun des termes possède deux syntagmes nominaux dont le syntagme de premier niveau est le même, **a sociedade**, on arrive aux termes **informatização na sociedade** et **informatização da sociedade** à partir du centre de syntagme nominal de premier niveau **sociedade**. Le même résultat peut être trouvé si l'on fait la recherche à partir du centre de syntagme nominal de deuxième niveau **informatização**. On observe donc, dans cette approche, que les changements de prépositions dans les termes ne nuisent pas aux résultats d'une recherche. M. LE GUERN montre dans son article, de la revue Le Français Moderne [59] , un exemple identique dans la langue française. Nous arrivons à la même constatation que lui ;
- **periódicos técnicos** (dans la maquette) au lieu de **periódicos técnico-científico** (dans le TCI) Le terme trouvé dans la maquette ressemble au terme du thesaurus. Il pourrait être classé comme un terme plus spécifique ou même un terme associé à celui du thesaurus.

D'autre part, il y a quelques syntagmes nominaux trouvés et qui se rattachent partiellement aux termes du TCI, comme :

- literatura economica
- literatura tecnica disponível
- literatura técnico-científica mundial
- tandis que dans le TCI on trouve les termes :
- literatura
- literatura de cordel (terme brésilien, c'est un type de littérature caractéristique du Nord-Est du Brésil qu'on trouve dans des feuilles pauvrement imprimées et qui sont placés pendus en cordeau dans les marchés et foires)
- literatura de ficção

[59] Michel LE GUERN, « Un analyseur morpho-syntaxique pour l'indexation automatique », *Le Français Moderne*. Juin, 1991, t. LIX, n°. 1, p. 34-35.

· literatura infantil

Le terme **literatura** dans la langue portugaise, aussi bien que **littérature** dans la langue française fait référence aux œuvres littéraires. Selon le dictionnaire Le Robert Micro, ***« ce sont des œuvres écrites, dans la mesure où elles portent la marque de préoccupations esthétiques ; les connaissances, les activités qui s'y rapportent »*** . Ainsi, il semble que les termes trouvés dans le TCI ont été regroupés en tenant compte de ce concept. Par contre, les termes **literatura economica**, **literatura técnica disponível** et **literatura técnico-científica mundial**, sont des termes plus récents qui ajoutent un deuxième concept au premier, tant au Brésil qu'en France : *bibliographie sur une question donnée*. Il s'agit d'une actualisation sémantique de ce terme.

Ces observations démontrent que les syntagmes nominaux extraits directement des textes composants d'une base de données, peut répondre plus précisément à la demande d'un utilisateur. En plus, ces observations peuvent suggérer l'utilisation de l'approche des syntagmes nominaux pour améliorer les thesaurus autant en ce qui concerne la qualification d'un terme qu'en ce qui concerne l'amélioration de la forme du terme.

3 Conclusion

D'une manière rapide et dans un contexte très restreint on a pu observer quelques faits qui montrent l'utilité des syntagmes nominaux comme moyen d'accès à l'information et qui peuvent ouvrir d'autres champs d'applications. En ce qui concerne les chiffres présentés, on peut les considérer comme raisonnables, étant donné : a) la petite taille du corpus ; b) le fait que les articles n'ont pas couvert toutes les catégories du TCI ; et c) le fait qu'il s'agit d'un thesaurus développé en 1989 (donc non actualisé).

La manque d'actualisation du thesaurus utilisé n'empêche pas de conclure que les syntagmes nominaux permettent obtenir des index de descripteurs plus actualisés que celui créé à partir d'un thesaurus. C'est une conséquence du fait que la mise à jour d'un index de syntagmes nominaux ne dépend que des articles car ils peuvent en être extraits directement et de manière automatique. Par contre, la mise à jour du thesaurus dépend de l'intervention humaine et de la saisie des termes candidats dans les ouvrages spécialisés.

Dans ce contexte le pourcentage de 10,76% des termes du TCI trouvé dans la maquette constitue un bon résultat. Il faut rendre compte des syntagmes nominaux existant dans la maquette et qui pourraient se rattacher au TCI, ce sont des termes vraisemblablement candidats au thesaurus. En tenant compte de ces aspects là, il paraît possible, que ce pourcentage puisse croître.

En outre, une autre conclusion se dégage : les syntagmes nominaux peuvent aider de manière plus efficace la construction et la mise à jour d'un thesaurus.

En ce qui concerne la maquette, la seule chose gênante est, lorsqu'on cherche un syntagme nominal de plus haut niveau, d'être obligé de commencer la recherche à partir du centre de syntagme nominal de premier niveau. Il faut donner aux utilisateurs la

possibilité d’accéder aux informations de manière plus rapide, en utilisant des centres de syntagmes nominaux de plus haut niveau. Ce genre de recherche est davantage destiné aux utilisateurs spécialistes alors que la recherche à partir du centre de syntagmes nominal de premier niveau est plus facile d’utilisation pour les débutants. Une autre solution possible pour satisfaire les usagers expérimentés est de leur permettre de faire la recherche directement avec le syntagme nominal complet, sans utiliser la navigation à partir d'un centre de syntagme nominal.

Troisième partie : le modèle de reconnaissance du SN

« La langue est une raison humaine qui a ses raisons, et que l'homme ne connaît pas. » Lévi-Strauss (Claude), La Pensée sauvage (Plon).

Chapitre 7 L'omission de déterminants dans le discours en langue portugaise

1 Considérations préliminaires

Le déterminant est un élément important dans un syntagme nominal (SN), car il est une marque du début de celui-ci. Cependant, dans la langue portugaise on trouve souvent dans le discours des SN sans déterminant, ce qui peut rendre difficile la tâche de reconnaissance et d'extraction automatique de ces derniers.

Dans la procédure d'extraction des SN mise en place dans le cadre de notre mémoire 60 de DEA, nous avons trouvé une incidence relativement grande d'absence de déterminant dans les SN extraits. Une statistique descriptive montre cet aspect dans le

tableau de la figure 7.1.

Syntagmes Nominaux	Quantité Absolue	Quantité Relative
Syntagmes Nominaux	6010	100,00%
SN avec déterminant Article défini article indéfini autres déterminants	4076 3586 490 198	71,11% 59,67% 8,15% 3,29%
SN sans déterminant	1736	28,89%

Les déterminants non-articles, dans la figure 7.1, correspondent aux adjectifs démonstratifs et indéfinis, aux numéraux cardinaux et aux chiffres. On constate donc, que la quantité de SN sans article et sans déterminant est assez important, soit plus d'un quart des SN.

Selon M. LE GUERN, le Français du seizième siècle, à la différence de celui d'aujourd'hui, admettait l'absence de l'article dans plusieurs contextes. CUNHA & CINTRA font référence à ce phénomène dans la phase primitive des langues romanes :

« A la rigueur, il ne s'agit pas proprement, dans ce cas et dans les suivants d'omission d'article indéfini, mais de cas où il n'a jamais été employé de manière régulière. « Dans la phase primitive des langues romanes, l'article indéfini était d'utilisation restreinte. Avec le temps, ce déterminatif s'est introduit dans plusieurs constructions et, aujourd'hui, les diverses nuances de son emploi constituent une inestimable richesse stylistique pour chacune d'entre elles. « Par contre, nos grammairiens refusent cette généralisation et valorisation progressive de l'article indéfini, où ils ne voient qu'une simple et superflue influence du français, et où, en fait, les interdictions d'utilisation de ce déterminatif sont actuellement peu nombreuses. Mais une telle discussion s'est révélée inutile parce qu'il ne s'agit pas d'un simple gallicisme qu'on peut extirper, mais d'une tendance générale des idiomes latins en recherche de formes plus expressives, avec plus de clarté et fermeté pour l'énoncé. » [61]

Cette observation concernant la langue portugaise confirme encore, par opposition à la

[60] Hélio KURAMOTO. *Maquette d'un système de recherche d'information en utilisant des syntagmes nominaux.* Villeurbanne, 1995. Mémoire du DEA. École Nationale Supérieure des Sciences de l'Information et des Bibliothèques.

[61] ***Celso CUNHA et Lindley CINTRA. Nova Gramática do Português Contemporâneo. Lisboa : Edições João Sá da Costa, 1991. p. 242 « Em rigor, não se trata propriamente, nesses casos e nos seguintes de omissão do artigo indefinido, mas de casos onde ele nunca se empregou de forma regular. « Na fase primitiva das línguas românicas, o artigo indefinido era de uso restrito. Com o correr do tempo, esse determinativo foi-se introduzindo em numerosas construções e, hoje, os variados matizes do seu emprego constituem uma inestimável riqueza estilística de todas elas. « Contra essa generalização e valorização progressiva do indefinido se manifestaram sempre os nossos gramáticos, que nela vêem uma simples e desnecessária influência do francês, onde, em verdade, poucas são actualmente as interdições ao uso do determinativo em causa. Mas tal guerra tem-se revelado inútil, e inútil precisamente porque não se trata, no caso, de um mero galicismo extirpável, e sim de uma tendência geral dos idiomas neolatinos em busca de formas mais expressivas, de maior clareza e vigor para o enunciado. »***

langue française, qu'il n'y a eu aucune évolution dans le sens de l'introduction du déterminatif indéfini. D'ailleurs, l'absence d'article défini et l'inexistence d'article partitif sont aussi des faits importants dans la langue portugaise. Ce qui corrobore un tel pourcentage de SN sans articles observés dans ce corpus.

Il faut donc étudier les contextes où les articles peuvent être omis et établir un ensemble de règles de façon à pouvoir extraire automatiquement les SN de manière plus précise.

2 Contextes d'omission d'articles dans la langue portugaise

Dans le discours en langue portugaise, l'omission d'article peut se produire dans plusieurs contextes. Parmi eux, deux cas sont tout à fait justifiables au niveau d'un SN, ce sont :

1. l'omission de l'article après une préposition et devant un nom ;
2. l'omission de l'article devant SN ayant un adjectif démonstratif, possessif, indéfini, ou numéral comme déterminant.

Dans le premier cas, il s'agit d'une expansion prépositionnelle [62] . Soit l'exemple suivant : *os móveis de cozinha* (les meubles de cuisine)

Dans cet exemple, le mot *cozinha* (cuisine) ne désigne qu'un ensemble de prédicats, il n'est lié à aucune réalité extra-linguistique. Il est un signe sans référence. On n'a pas ici un cas d'omission d'article mais un cas où l'article ne peut pas être employé.

En ce qui concerne le deuxième cas (2) plus haut, il s'agit du fait que l'adjectif démonstratif, possessif ou indéfini fonctionne déjà comme un déterminant, ce qui justifie l'absence d'article.

Exemple : *Este sistema de informação* (Ce système d'information)

Ainsi on passe à énumérer les autres cas d'omission d'article défini, selon Celso CUNHA & Lindley CINTRA, dans le discours en langue portugaise :

1. Devant un substantif abstrait exprimant la totalité d'un genre, d'une catégorie, d'un groupe, d'une substance, ou lorsqu'il fait partie des proverbes, des phrases sentencieuses, et des comparaisons brèves. Exemple : Pobreza não é vilezaCe cas rappelle le Français du XVIème siècle où l'article était peu utilisé. Au XVIIe siècle l'utilisation d'article était plus courant. Pourtant, on l'omettait devant les noms abstraits, tels que : amour, nature, fortune, mort, etc., sans qu'ils soient toujours à proprement personnifiés. (voir HAASE) Exemple : Ma présence importune Te laisse à la merci d'Amour et de la brune. (Corn., Mél., I, 5, 346.)(La pauvreté n'est pas vice) Homem não é bicho (L'homme n'est pas une bête).

[62] L'opposition entre le syntagme prépositionnel (SP) et l'expansion prépositionnelle (EP) se traduit par le fait qu'un N' peut dominer un N'', ce qui est impossible pour un N. (SP □ P' + N'' Exemple: Le placard de la cuisine … SP □ de + la + cuisine ; tandis que EP □ P' + N Exemple : Le placard de cuisine… EP □ de + cuisine). Voir Michel LE GUERN. « Un analyseur morpho-syntaxique pour l'indexation automatique ». *Le Français Moderne*, juin, 1991, p. 29.

2. Comme en français les noms de mois n'admettent pas l'article, sauf quand ils sont suivis d'un qualificatif. Exemples : Estou seguro de ir até o Rio em fins de junho ou princípios de Julho (Je suis sûr d'aller à Rio à la fin de juin ou au début de juillet). (Mário de Andrade, CMB, 102.) ; Era um setembro negro (C'était un septembre noir) ;

3. On omet l'article devant les dates du mois. Exemple : O parecer é de 28 de janeiro de 1640 (l'opinion est du 28 janvier 1640). Cependant l'article est utilisé lorsque la date est célèbre, lequel acquis la valeur d'un substantif composé. Exemple : Por ser precisamente um dos feriados extintos o 19 de Novembro faz lembrar hoje... (Pour être précisément un des jours fériés supprimés le 19 novembre fait rappeler aujourd'hui...). (Carlos Drumond de Andrade, FA, 116.). On utilise l'article devant une date mentionnée dans le cours d'une narration. Exemple : Constituiu-se assim livremente a Academia e a primeira sessão se realizou aos 15 de dezembro de 1896... (Ainsi fut librement constitué l'Académie et la première session eut lieu le 15 décembre 1896...). (Manuel Bandeira, PP, II, 1132.) ;

4. Comme en français on n'utilise pas l'article devant les heures du jour, ni avec des expressions comme meio-dia (midi) et meia-noite (mi-nuit). Exemple : O relógio marcava meio-dia e dez...(l'horloge indiquait minuit et dix minutes). Cependant, on l'utilise lorsqu'ils sont précédés de prépositions. Exemple : Já não se almoça às 9 da manhã e não se janta às 4 (On ne déjeune plus à 9 heures du matin et on ne dîne pas à 4 heure de l'après midi). (Carlos Drumond de Andrade, MA, 99.) ;

5. D'une manière générale, on utilise l'article devant les noms de pays, de régions, de continents, de montagnes, de volcans, de déserts, de constellations, de fleuves, de lacs, d'océans, de mers et de groupes d'îles. Cependant, il y a quelques noms de pays et de régions qui refusent l'article, comme : Portugal, Angola, Moçambique, Cabo Verde, São Tomé, Príncipe, Macau, Timor, Andorra, Israel, São Salvador, Aragão, Castela, Leão ;

6. Les noms de villes, de lieux, de planètes, d'étoiles et de la plus grande partie des îles, d'une façon générale, ne sont pas précédés d'article. Cependant, on utilise l'article au cas où le nom de ville est formé d'un substantif commun. Exemple : o Porto, o Rio de Janeiro, a Guarda, o Cairo, a Haia ;

7. Après le mot todo (tout), au Brésil principalement, on utilise l'article pour distinguer entre le sens tout « quelconque », « chaque » du sens tout « entier », « total ». Exemple : Toda casa [=uma casa de uma forma geral] cedo ou tarde precisa de reforma (Toute maison [=maison quelconque] tôt ou tard a besoin de rénovation) — Toda a casa [=a casa inteira] foi reformada (Toute la maison [=la maison entière] était rénovée. En ce cas, les mots toda, todas, todo, todos font partie de la catégorie de prédétermiants de notre grammaire de référence (voir les prédéterminants dans le chapitre 8) ;

8. On omet l'article dans les énumérations lorsqu'on veut obtenir un effet d'accumulation ou de dispersion. Exemple : perspectivas do agente da informação no contexto brasileiro: problemas, barreiras e desafios (perspectives de l'agent de l'information dans le contexte brésilien : problèmes, barrières, défis) ;

9. Comme en français, on répète l'article en portugais devant les substantifs coordonnés, sauf lorsqu'ils représentent un tout strictement uni. En ce cas, on ne met l'article que devant le premier substantif avec lequel il s'accorde en genre et en nombre. Exemples : O estudo [do folclore] era necessitado pela existencia das historias, contos de fadas, fabulas, apologos, superstições, provérbios, poesia e mitos recolhidos da tradiçao oral (L'étude du folklore était rendue nécessaire par l'existence des histoires, des contes de fées; des fables, des apologues, des superstitions, des proverbes, des poésies et des mythes issus de la tradition orale). (Joao Ribeiro, FI, 6.) — O engenho e arte (Le génie et l'art) — Os direitos e deveres (Les droits et devoirs) ;

10. On omet l'article dans les vocatifs. Exemple : Oh! dias da minha infância! Oh! meu céu de primavera! (Oh! jours de mon enfance! Oh! mon ciel de printemps!) ;

11. On omet l'article dans l'apposition qui indique une simple appréciation. Exemple : Tardes de minha terra, doce encanto, Tardes duma pureza de açucenas (Soirées de ma terre, doux enchantement, (Soirées d'une pureté d'amaryllisAmaryllis est, selon le dictionnaire LAROUSSE, en CD-ROM, une plante de la famille des amaryllidacée la même que les « açucenas », dans laquelle appartient aussi la Fleur de Lis, le Lis Saint-Jacques.) (Florbela Espanca, S, 35.) ;

12. Devant les mots qui désignent les matières d'études, employés avec des verbes comme : aprender (apprendre), estudar (étudier), cursar (suivre cours), ensinar (enseigner) et synonymes. Exemples : Aprender Inglês (Apprendre l'Anglais) — Estudar Latin (Étudier le Latin) — Ensinar Geometria (Enseigner la Géométrie) ;

13. On omet l'article devant les mots tempo (temps), ocasião (occasion), motivo (motif), permissão (permission), força (force), valor (valeur), ânimo (esprit), compléments des verbes ter (avoir), dar (donner), pedir (demander) et ses synonymes. Exemples : Não houve tempo para descanso (Il n'y a pas de temps pour se reposer) — Não dei motivo à crítica (Je n'ai pas donné motif à la critique) ;

14. Le nom propre par définition n'oblige pas l'utilisation d'article car il est déjà déterminé. Or, au cours de l'histoire de la langue, pour diverses raisons cette norme logique n'est pas toujours observée. Aujourd'hui, il y a plusieurs situations où le nom propre oblige l'utilisation de l'article : soit pour renforcer une idée d'individualisme, soit parce qu'il y a quelques noms propres qui proviennent des noms communs, soit à cause de l'influence de la langue italienne où le nom de famille quand il est utilisé tout seul vient précédé de l'article ou encore pour donner une atmosphère familière.

On omet l'article indéfini dans les contextes suivants :

1. lorsqu'on a un autre élément déterminatif avant le nom, par exemple, une forme d'identité ou de comparaison. Exemple : De você não esperava semelhante gesto (Je n'attendais pas de toi un tel geste) ;

2. quand on emploi un substantif pour désigner toute une espèce ou catégorie. Exemple : Amigo fiel e prudente é melhor que parente (Un ami fidèle et prudent est meilleur qu'un parent) ;

3. On évite l'article indéfini quand il y a déjà devant le substantif un des adjectifs démonstratifs igual, semelhante et tal ; ou undes pronoms indéfinis certo, qualquer, outro et tanto. Exemple : Certo amigo meu já usou de igual argumento (Un de mes amis a déjà utilisé pareil argument). Cependant lorsqu'on utilise quelques unes de ces formes postposées à un substantif, elles jouent le rôle d'adjectifs. En ce cas il est normal que le substantif soit acompagné par un article indéfini Exemple : Quero um livro igual a esse (Je veux un livre comme ça.). On omet l'article quand la phrase est négative ou interrogative. Exemple : Nunca li coisa igual (Je n'ai jamais lu une telle chose) ;

4. En principe, les formules comparatives peuvent admettre l'exclusion de l'article. Exemple : Nunca passei por lugar tão perigoso como aquele (Je ne suis jamais passé par un endroit aussi dangereux que celui-là) — Trabalhava com tanto cuidado como o pai (Il travaillait aussi soigneusement que le père) — Não encontrarias melhor amigo nesta emergência (Vous ne trouveriez pas meilleur ami en cas de besoin) — Que furacão, revolveu tudo (Quelle tornade, elle a tout bouleversé) ;

5. Il est habituel d'éviter l'article indéfini devant des expressions dénotant des quantités indéterminées, constituées soit par des substantifs (comme : coisa, gente, infinidade, número, parte, pessoa, porção, quantia, quantidade, soma et equivalents), soit par des adjectifs (comme : escasso, excessivo, suficiente et synonymes). Exemple : Havia grande número de pessoas no casamento (Il y avait un grand nombre de personnes au mariage). En fait ces noms (adjectifs et substantifs) font partie des mots qui peuvent entrer dans la formation des déterminants complexesOn entend comme déterminant complexe, ceux composés par article, nom substantif et adjectifs, numéral cardinal comme dans : les deux, les plus grands, un demi kilo de, trois de, etc. puisqu'ils désignent une quantification encore qu'indéfinie ;

6. La présence du numéral meio empêche d'une manière générale l'article indéfini. Exemple : Comprou meio quilo de pão (Il a acheté un demi-kilo de pain) — Tomou meia dose do remédio (Il a pris une demie dose du médicament). Cependant, le mot féminin meia, construit avec l'article indéfini donne une désignation de quantité approximative, ou quand il forme avec le substantif une unité d'utilisation courante. Exemple : Só tenho uma meia libra (Je n'ai qu'une demie livre) — No caso, basta uma meia-palavra sua (En ce cas, il suffit d'un demi-mot de lui). Là encore, on peut faire la même observation que dans l'item précédent ;

7. Comme pour les articles définis, on omet l'article indéfini dans les énumérations et appositions. Exemple : Casas, árvores, nuvens desagregavam-se numa melancólica paisagem de Outono (Des maisons, des arbres, des nuages se désagrégeaient dans un mélancolique paysage d'automne). (Fernando Namora, TJ, 232.) ;

8. Quand un substantif au singulier est conçu sur l'aspect de catégorie, d'espèce et non pas sur l'aspect d'unité, l'article indéfini être omis. Exemple : Cão ladrador nunca é bom caçador (Un chien qui aboie n'est jamais un bon chasseur).

D'ailleurs, on trouve d'autres remarques faites par Paul TEYSSIER [63] comme :

[63] Paul TEYSSIER. *Manuel de Langue Portugaise : Portugal-Brésil*. Paris : Editions Klincksieck, 1984, p. 78-79.

- on n'emploie pas l'article défini dans des locutions formées d'une préposition suivie d'un substantif. Exemple : *em nome de* (au nom de) — *em benefício de* (en bénéfice de) — *para uso de* (pour usage de) — *por conselho de* (sur le conseil de). Dans ce cas on n'omet pas volontairement l'article, celui-ci n'est jamais utilisé : ce sont des locutions prépositionnelles ;
- L'omission de l'article défini se trouve aussi dans des expressions figées formées d'un verbe suivi d'un substantif. Exemple : *sentir necessidade de* (éprouver le besoin de) — *é costume* (c'est une habitude) — *prestar atenção a* (faire attention à) — *ter direito a* (avoir droit à). En fait, on observe là, le même phénomène observé par A. HAASE et rappelé par M. LE GUERN, sur l'absence d'articles devant les noms abstraits dans des constructions de phrases du français au seizième siècle. Ce qui demeure dans la langue portugaise. Ainsi, ce n'est pas le fait qu'il s'agit ici d'une expression figée, mais principalement des noms abstraits.

En ce qui concerne l'article indéfini, TEYSSIER remarque que son utilisation est une marque de *vernaculidade*, c'est-à-dire une marque d'originalité, celle de vouloir parler la vraie langue portugaise. Cependant, l'abus de l'usage d'article indéfini apparaît parfois comme une imitation consciente ou inconsciente du français.

D'autre part, dans le portugais moderne, il n'y a rien qui ressemble à l'article partitif français. Le partitif français se rend, en Portugais, simplement par le substantif sans article. Exemple : du vin = *vinho* — de l'eau = *água* — du pain = *pão*, etc.

3 Analyse de quelques syntagmes nominaux avec déterminant zéro

La section précédente a montré quelques contextes où l'omission se fait de manière régulière. Or, l'on trouve d'autres contextes qui ne suivent pas des règles régulières. Ce sont en fait des exceptions à une règle donnée, soit parce que des mots refusent l'article, soit parce que l'absence d'article donne un sens différent à la phrase. En plus, il faut remarquer qu'il n'y a pas d'article partitif dans la langue portugaise et que rien ne le remplace. D'ailleurs, il semble qu'une partie des contextes présentés apparaissent plutôt dans les ouvrages littéraires que dans les textes techniques. Il faut savoir dans quels contextes (textes techniques) se trouve l'article zéro. Ainsi, l'analyse des syntagmes nominaux extraits d'un corpus de textes techniques est nécessaire. Pour cela, nous allons reprendre les syntagmes nominaux extraits, dans le cadre du DEA, d'un corpus de 15 (quinze) articles du domaine de sciences de l'information.

Parmi les 1736 SN sans articles nous en avons choisi quelques-uns uns pour les analyser à la lumière des contextes présentés plus haut. Les SN analysés sont regroupés selon quelques caractéristiques que nous allons montrer.

Dans un syntagme prépositionnel 1.

Textes en portugais	Textes en français
1. O surgimento de uma categoria de clientes conscientes dos seus direitos a produtos e serviços de alta qualidade.	L'apparition d'une catégorie de clients conscients de leurs droits à des produits et à des services de haute qualité.
2. O movimento de mudanças em direção à melhor sintonia com o mercado.	Le mouvement de changements vers une meilleure syntonie avec le marché.
3. ... o conjunto de estratégias.	... l'ensemble de stratégies.
4. ... o alcance de objetivos préestabelecidos	... l'atteinte d'objectifs préétablis.
5. ... o conjunto formado por recursos humanos capacitados	... l'ensemble formé par des ressources humaines entraînées
6. ... o uso de clientes específicos.	... l'usage de clients spécifiques.
7. ... produzindo através da experiência profissional refletida, do treinamento contextualizado, de manuais de rotinas, de sistemas especialistas.	... produisant au moyen de l'expérience professionnelle réfléchie, d'entraînement contextualisé, de manuels de rotines, de systèmes spécialistes.
8. ... ensinar a futuros empregados.	... enseigner aux futurs employés.
9. ... destinados a áreas e setores específicos	... destinés aux domaines et secteurs spécifiques.
10. ... o gerente de recursos informacionais	... l'administrateur de ressources d'information.
11. ... conduzir decisões participativas sobre tecnologias e sistemas de informação	... conduire des décisions participatives sur des technologies et sur des systèmes d'information.
12. cientistas envolvidos em pesquisas biologicas, psicologicas ou sociais	des scientifiques engagés dans des recherches biologiques, psychologiques ou sociaux
13. Sistemas de apoio à decisão, baseados em inteligência artificial e modelos matemáticos da realidade	Les systèmes d'aide à la décision, basés sur l'intelligence artificielle et sur les modèles mathématiques de la réalité
14. Esse saber codificado sob forma de informações e sistemas.	Ce savoir codé sous forme d'informations et des systèmes.

On se rend compte, dans ces exemples, que l'inexistence d'article partitif dans la langue portugaise entraîne l'omission d'article. Ce qui peut rendre difficile l'identification d'un SN. En revanche, bien que l'article soit omis, il est possible d'observer que les termes suivant la préposition sont soit au pluriel, soit abstraits.

Après un verbe 1.

Textes en portugais	Textes en français
1. Um elenco de recursos estratégicos capazes de propiciar vantagem competitiva diante da concorrência às organizações-líderes	Un ensemble de ressources stratégiques capables de favoriser l'avantage compétitif devant la concurrence aux organisations-leaders.
2. Tais sistemas incorporam estruturas informacionais, tecnológicas e educacionais	De tels systèmes regroupent des structures d'information, technologiques et éducationnelles
3. Esses sistemas associam capacidade de processamento convencional com habilidade lógica de solução de problemas e de aconselhamento especialista	Ces systèmes associent la capacité de traitement conventionnel avec l'habileté logique de solution de problèmes et d'orientation spécialiste
4. Criar condições internas para transformação da informação e da tecnologia em qualidade, produtividade e lucro	Créer des conditions internes pour la transformation de l'information et de la technologie en qualité, productivité et profit
5. Centros de análise de informação localizam conteúdos de diferentes fontes, que são analisadas e sintetizadas sob forma de novos produtos de alto valor agregado	Des centres d'analyse de l'information localisent des contenus de différentes ressources, qui sont analysées et synthétisées sous forme de nouveaux produits de haute valeur ajoutée
6. Uma expressão que designa sistematizações relacionadas com as evoluções e mutações que marcaram a economia dos países desenvolvidos	Une expression qui désigne des systématisations associées avec les évolutions et les mutations qui ont marqué l'économie des pays développés
7. O estudo da eficácia compara objetivos e resultados	L'étude de l'efficacité compare des objectifs et des résultats
8. A importante finalidade de identificar capacitações na empresa que possam ajudar ou prejudicar o aproveitamento das oportunidades	L'importante finalité d'identifier des formations dans l'entreprise qui puissent aider ou nuire au profit des opportunités
9. Toda informação pode trazer benefícios	Toute information peut apporter des bénéfices
10. ...deve apresentar características bem diversas da informação que é usada para a gestão empresarial dessa mesma empresa	... doit présenter des caractéristiques très diverses de l'information utilisée pour la gestion des affaires de cette même entreprise
11. desenvolver politicas, procedimentos, diretrizes e sistematicas para a organização da função	développer des politiques, des procédures, des directives et des systématiques pour l'organisation de la fonction

En ce cas, on observe l'omission d'article devant les termes qui apparaissent après

un verbe à l'infinitif ou fléchi. Ici encore on voit l'absence d'article comme une conséquence de l'inexistence de l'article partitif dans la langue portugaise. On constate là que les termes sont souvent soit au pluriel soit abstraits.

Dans une suite de termes coordonnés — type 1 (énumération) 1.

Textes en portugais	Textes en français
1. ... mediante a educação formal, treinamento e comunicação	... selon l'éducation formelle, l'entraînement et la communication
2. Assim, a produção, difusão, assimilação e uso estratégico do conhecimento empresarial	Ainsi, la production, la diffusion, l'assimilation et l'usage stratégique de la connaissance des affaires
3. A análise, interpretação, avaliação e comunicação da informação pelos meios convenientes	L'analyse, l'interprétation, l'évaluation et la communication de l'information par les moyens convenables
4. as funções dos responsaveis pela implantação, manutenção e aperfeicoamento das unidades de informação	les fonctions des responsables pour l'implantation, l'entretien et le perfectionnement des unités d'information
5. a classificação, organização e recuperação de informações	la classification, l'organisation et la recherche d'informations
6. a armazenagem, recuperação e utilização de documentos / infor-mações nas organizações	le stockage, la recherche et l'utilisa-tion des documents / informations dans les organisations
7. as contribuições, desempenhos e funções das entidades...	les contributions, les performances et les fonctions des entités ...
8. levantamento do fluxo, volume e taxa de atualização dos dados	la recherche du flux, du volume et du taux d'actualisation des données
9. caracterização dos encargos, deveres e responsabilidades na execução de tarefas e rotinas	caractérisation des charges, des devoirs et des responsabilités dans l'exécution de tâches et routines
10. uma utilização, acessibilidade e disseminação mais eficientes	une utilisation, une accessibilité et une dissémination plus performantes
11. o projeto, estruturação e disseminação de informações	le projet, la structuration et la dissémination d'informations
12. o pessoal empregado nas pequenas e medias industrias	les personnes employés dans les petites et moyennes industries

Ce cas a été déjà prévu dans les contextes d'omission d'articles présentés dans la section précédente. Couramment on répète l'article devant les substantifs coordonnés, sauf lorsqu'ils représentent un tout strictement uni. En ce cas, on ne met l'article que devant le premier substantif avec lequel il s'accorde en genre et en nombre. On rend compte donc qu'il s'agit d'une réelle omission d'article. Ainsi, rien n'empêche d'adopter la solution de mettre, dans ce contexte, l'article correspondant où l'article est absent.

Dans une suite de termes coordonnés — type 2 (énumération) 1.

Textes en portugais	Textes en français
1. Liberdade, maturidade, experiência, comunicação, sensibilidade, criatividade, intuição e educação continuada constituem...	La liberté, la maturité, l'expérience, la communication, la sensibilité, la créativité, l'intuition et l'éducation continue forment...

Cet exemple ressemble à ceux du contexte 3 dont la seule différence est l'omission totale de l'article. D'autre part, nous voyons que les substantifs sont abstraits. Ce qui corrobore, encore une fois, l'omission de l'article devant les noms abstraits.

Enumération sans article — type 3 1.

Textes en portugais	Textes en français
1. formação de redes (alianças, jointventures, consorcios, etc)	formation de réseau (alliances, jointventures, consortiums, etc.)
2. Seu estado constante em cruzeiro (altura, velocidade e rota)	Son état constant en croisière (hauteur, vitesse et route)
3. a criação de novas carreiras na profissão do bibliotecário, como : especialista da informação, agente da informação, profissional da informação, cientista da informação, administrador de recursos informacionais e outros.	la création de nouvelles carrières dans la profession de bibliothécaire, comme : spécialiste de l'informa-tion, agent de l'information, pro-fessionnel de l'information, scientifique de l'information, administrateur de ressources informationnelles et d'autres

On trouve encore d'autres cas d'énumération avec omission d'articles. Comme nous l'avons déjà remarqué, on peut toujours repérer les syntagmes nominaux en observant les mots qui sont dans l'énumération, plus spécifiquement s'ils sont au pluriel, s'ils sont des noms abstraits ou s'ils ont un syntagme prépositionnel. En fait, on cherche à repérer si les mots forment un prédicat lié. Or, il faut faire attention parce qu'on peut trouver des énumérations où les mots ne désignent que des simples prédicats, sans référence à un univers donné. Le cas numéro 2 du tableau plus haut constitue un exemple de ce contexte. Ils ne sont qu'une suite de mots qui sont plutôt des propriétés. Il ne s'agit donc pas de SN. On risque d'extraire des SN parasites.

Début de paragraphe 1.

Textes en portugais	Textes en français
1. Conhecimento humano especializado é o...	La connaissance humaine spécialisée est le ...
2. Bibliotecas e centros de informação facilitam...	Les bibliothèques et les centres d'information permettent
3. Centros de análise de informação localizam ...	Les centres d'analyse d'information localisent ...
4. modelos de avaliação de impactos em analise prospectiva	Des modèles d'évaluation des impacts sur l'analyse prospective
5. comercialização, divulgação e marketing dos serviços de informação	la commercialisation, la divulgation et le marketing des services d'information
6. Instrumentos que caracterizam os fatores de vantagem competitiva	Des instruments qui caractérisent les facteurs des avantage
7. Perguntas que interessam à economia da informação	Des questions qui intéressent à l'économie d'information
8. produtos de informação tem ...	les produits d'information ont ...
9. Sistemas de apoio à decisão, baseados em inteligência artificial e modelos matemáticos da realidade	Les systèmes d'aide à la décision, basés sur l'intelligence artificielle et sur les modèles mathématiques de la réalité
10. Centros de informação e bancos de dados coletam...	Les centres d'information et banques de données prennent ...

Bien que nous avons classé ce type d'omission comme étant des SN placés en début de paragraphe, ceci n'est pas vrai. Nous avons regardé plusieurs paragraphes dans une dizaine d'exemplaires de la revue Revista Ciência da Informação et nous sommes arrivés à la conclusion de que ce type d'omission ne constitue pas une règle ou un style d'écriture. Or, si nous analysons chacun des SN présentés, nous pouvons se rendre compte que les mots initiaux de chaque SN sont soit au pluriel, soit abstraits. Voyons les deux premiers exemples, nous pouvons dire qu'ils sont des SN. Le premier parce que le mot *conhecimento* (connaissance) est un nom abstrait. Dans le deuxième exemple, il s'agit d'une suite de mots coordonnés où le mot *bibliotecas* est au pluriel, donc, un prédicat lié. Dans le deuxième terme de cette suite coordonnée, *centros de informação*, il s'agit aussi d'un prédicat lié car on a là le mot centros au pluriel.

1. L'absence d'article dans l'apposition

Selon Celso CUNHA & Lindley CINTRA [64] , on omet l'article défini dans l'apposition lorsqu'elle indique des simples appréciations. C'est-à-dire, il y a des cas d'apposition qui sont précédés d'un déterminant. Il nos semble qu'en cas d'omission d'article nous pouvons utiliser les mêmes repères des autres cas d'omission de déterminant : voir si le nom est au pluriel ou s'il est un nom abstrait. En effet, nous proposons ici de la reconnaître comme étant un nouveau syntagme nominal.

[64] Celson CUNHA & Lindley CINTRA. *Nova Gramatica do Português Contemporâneo.* Lisboa : Edições João de Sá da Costa Ltd., 1984. p. 238.

Textes en portugais	Textes en français
1. incorporam o espaço, ou ambiente circundante.	incorporent l'espace, ou environnement tournant.
2. a validade dos fatores de diagnóstico, proposição e satisfação, considerados na execução do serviço	la validité des facteurs de diagnosti-que, proposition et satisfaction, con-siderés dans l'exécution du service
3. A consultoria, enquanto fonte do conhecimento empresarial, ...	La consultation, en tant que source de la connaissance des affaires, ...

Dans une apparente expression prépositionnelle 1.

Textes en portugais	Textes en français
1. ...a produção do conhecimento mediante sistema de pesquisa e desenvolvimento	... la production de la connaissance au moyen d'un système de recherche et de développement
2. o especialista dotado de conhecimento teórico e experiência prática	le spécialiste doté de connaissance théorique et d'expérience pratique
3. as informações estratégicas para análise da concorrência	les informations stratégiques pour l'analyse de la concurrence
4. Esses sistemas associam capacidade de processamento convencional com habilidade lógica de solução de problemas e de aconselhamento especialista	Ces systèmes associent la capacité de traitement conventionnel avec la habileté logique de solution de problèmes et d'orientation spécialisée
5. as estruturas nacional e internacional de informação, documentação e de serviços de biblioteca	les structures nationales et internationales d'information, de documentation et de services de bibliothèque
6. transformação da informação em qualidade, produtividade e lucro	la transformation de l'information en qualité, productivité et profit

Les exemples présentés sont peut-être les plus difficiles dans la procédure d'extraction des SN. Mais, là encore les mots qui constituent les centres de SN sont des mots abstraits (système, connaissance, analyse, habileté, information, services, qualité, productivité, etc.). Or, dans l'exemple (1), après le syntagme « sistema… » (système…) il y a une deuxième expansion prépositionnelle, « de pesquisa e desenvolvimento » (« de recherche et de développement), et bien que le mot « pesquisa » soit un nom abstrait, il ne s'agit pas d'un SN. Tandis que dans l'exemple (2) on considère que la suite « *de coonhecimento teórico e experiência pratica* » (« de la connaissance théorique et de l'expérience pratique ») n'est pas une expansion prépositionnelle. Comment expliquer cette différence ? La principale différence entre les deux exemples est due au fait que le

mot « *sistema* », dans le premier exemple, n'exige pas un complément, tandis que le mot « *dotado* », dans le deuxième exemple après le mot « conhecimento », exige un complément. Il nous semble que cette différence explique pourquoi la suite du deuxième exemple ne peut pas être une expansion prépositionnelle.

Noms propres et sigles 1.

Les noms propres sont par définition des syntagmes nominaux, même sans article. Cependant, dans l'écriture en langue portugaise on trouve souvent ces noms précédés d'articles. Ce qui ne gêne pas l'identification de ce type de SN. Or, la question est de savoir comment identifier un nom propre car certains noms propres étaient à l'origine des noms communs. La seule chose qui peut aider est le fait que les noms propres commencent par une lettre majuscule. On peut donc les repérer plus facilement, les débuts de paragraphes, les titres commencent aussi par une lettre majuscule. En ce qui concerne les sigles, on peut les considérer comme étant un cas particulier de noms propres, donc ils peuvent être traités comme tels. De toute façon, les noms propres ainsi que les sigles serons mis dans le lexique et serons caractérisés ou classés, dans de catégories et sous-catégories, par un technicien.

Textes en portugais	Textes en français
1. ABC	ABC
2. Adriani	Adriani
3. Allen	Allen

Expressions entre guillemets 1.

Textes en portugais	Textes en français
1. « fluxos tensos »	« flux tendus »
2. « inteligência competitiva »	« intelligence compétitive »
3. « economia da informação »	« économie de l'information »

On trouve souvent des termes entre guillemets qui ont la configuration d'un prédicat lié, donc sans articles ou aucun déterminant. Ici, comme dans les cas d'énumération on peut repérer les termes à l'intérieur des guillemets comme des SN dès qu'ils font partie d'un prédicat lié. En ce cas nous pouvons les considérer comme étant des SN avec déterminant zéro.

Les contextes présentés possèdent diverses variations en ce qui concerne la distribution syntaxique. Ce qui a rendu beaucoup plus difficile la caractérisation de chaque contexte.

4 Conclusion

L'identification des SN avec article zéro ne peut pas être résolue seulement au moyen de l'observation du contexte syntaxique mais aussi grâce à un raisonnement logique. Nous

avons déjà fait quelques remarques là-dessus dans l'analyse des contextes présentés dans cette section.

L'idée principale de l'approche des syntagmes nominaux est que ceux-ci sont en relation avec la réalité extralinguistique de l'auteur en opposition aux mots clés. Les mots clés, souvent composés d'un simple mot de la langue, ne font pas référence aux objets du monde réel, ce sont des signes sans références. On a ici intervention de la logique intensionnelle où le signe ne désigne qu'un ensemble de prédicats. On parle alors du générique. En revanche, lorsqu'on parle du syntagme nominal on parle du spécifique. C'est l'intervention de la logique extensionnelle. Le syntagme nominal fait référence à la réalité extra-linguistique car il est un prédicat lié, déterminé soit par un article, soit par un autre type de déterminant (adjectif démonstratif, un pronom indéfini, etc.). Un prédicat lié est donc la transition entre la logique intensionnelle et la logique extensionnelle. Le prédicat lié fait référence, encore sans détermination, à une réalité extralinguistique.

Nous avons observé, dans les contextes des SN extraits du corpus de notre mémoire de DEA, que l'omission d'article se trouve souvent devant des prédicats liés parce que ceux-ci sont des termes déterminés, soit par un syntagme prépositionnel, soit par des noms abstraits. On trouve encore des termes qui sont au pluriel, ce qui permet de penser qu'il s'agit d'une référence à la réalité extralinguistique. On peut supposer que l'auteur, au moins dans sa pensée, fait référence à plus d'un objet du monde réel.

En ce qui concerne les noms abstraits, Marie-Noëlle GARY-PRIEUR fait une parallèle entre ceux-ci et les noms propres. Selon elle, ***« l'absence d'article caractérise on le sait l'usage le plus typique des noms propres. On a souvent relevé que les noms abstraits, eux aussi, ont plus longtemps que d'autres 'résisté' à l'article »*** [65] .

Cette affirmation, d'ailleurs signalée par M. LE GUERN, vient de corroborer ce que nous avons vu dans les exemples montrés ici.

Marie-Noëlle GARY-PRIEUR dit encore que tant les noms propres que les noms abstraits font renvoi à des entités uniques.

« Contrairement aux noms concrets, les noms propres et les noms abstraits à travers les variations d'un discours à l'autre, renvoient toujours au 'même référent'. On peut illustrer rapidement cette propriété par la comparaison des SN suivants :

a) la sagesse de Pierre la sagesse de Paul la sagesse de Jacques	c) le chien de Paul le chien de Zola le chien de Pierre
b) le Paris de Zola le Paris de Hugo le Paris de Maupassant	d) l'eau de la Seine l'eau de la mer l'eau de cette bouteille

Les SN (c) et (d) renvoient à des objets différents : ce n'est ni le même chien ni la même eau qui sont visés dans chacun des trois exemples. En (a) et (b), au contraire, c'est la même qualité (sagesse) et la même ville (Paris) qui sont visés à

[65] Marie-Noëlle GARY-PRIEUR. « A propos du fonctionnement sémantique des noms propres et des noms abstraits.». In. : Nelly Flaux, Michel GLATIGNY et Didier SAMAIN. *Les Noms Abstraits : histoire et théories.* Collection Sens et Structures. Paris : Presses universitaires du Septentrion, 1996, p. *136.*

travers différents aspects définis par le complément du nom. » [66] **.**

La procédure d'extraction des SN doit tenir compte des considérations faites ici. On peut suivre deux démarches pour la construction de cette procédure : 1) traitement des cas d'omission d'article en mettant un article où il faut ; et 2) extraction des syntagmes nominaux. Dans la première démarche, on fait une double analyse, tandis que dans la deuxième démarche, on résout l'omission d'article au fur et à mesure qu'on extrait les SN.

La deuxième démarche est plus acceptable car on fait l'analyse lexique une seule fois. Il reste, par contre, la question de remplacer l'article zéro par un vrai article. LE GUERN a proposé, dans son article sur la syntaxe des titres [67] , de mettre l'article devant les syntagmes nominaux avec article zéro dans les titres. Cette solution est très simple et elle a l'avantage de donner une homogénéité à l'ensemble des SN, sinon on aurait deux SN avec le même sens et de forme distincte. Or, il faut tenir compte des contextes des SN dans le discours car dans certains cas les prédicats liés sont ouverts et si on met un article défini, en fait, on procède à une fermeture logique. Il faut réfléchir sur cette nuance logique et les avantages de la solution proposée par LE GUERN. Avec l'homogénéité d'ensemble des SN obtenue à cause d'apposition d'article devant les SN, on a aussi une économie d'espace de stockage et une meilleure interaction avec l'utilisateur.

« Le désordre c'est l'ordre vu différemment. » HYREN , 283 av. JC

Chapitre 8 Proposition d'une démarche pour le développement du nouveau SRI

Nous allons montrer, dans ce chapitre, une séquence de procédures nécessaires au développement du nouveau SRI. Parmi ces procédures il y a l'élaboration d'un modèle pour l'identification et l'extraction automatique de syntagmes nominaux existant dans des textes en langue portugaise. En suite, il faut construire le système de reconnaissance et d'extraction de syntagmes nominaux et puis les modules d'indexation automatique ainsi que l'interface de recherche d'information.

Le modèle d'identification et d'extraction de syntagmes nominaux est composé, en fait, par deux modules distincts : 1) module de description lexicale ; et 2) module Grammaire de Reconnaissance et d'Extraction de syntagmes nominaux. Le module de description lexicale est présenté dans le prochain chapitre (chapitre 10) et le module Grammaire de Reconnaissance et d'Extraction de syntagmes nominaux est présenté

[66] ***Marie-Noëlle GARY-PRIEUR. « A propos du fonctionnement sémantique des noms propres et des noms abstraits.». In. : Nelly Flaux, Michel GLATIGNY et Didier SAMAIN. Les Noms Abstraits : histoire et théories. Collection Sens et Structures. Paris : Presses universitaires du Septentrion, 1996, p. 137-138.***

[67] Michel LE GUERN. « Traitement automatique et variation linguistique : la syntaxe des titres ». *Opérateurs et constructions syntaxiques : Evolutions des marques et des distributions du Xvème siècle.* Paris : Presses de l'Ecole Normale Supérieure, 1994, p. 79.

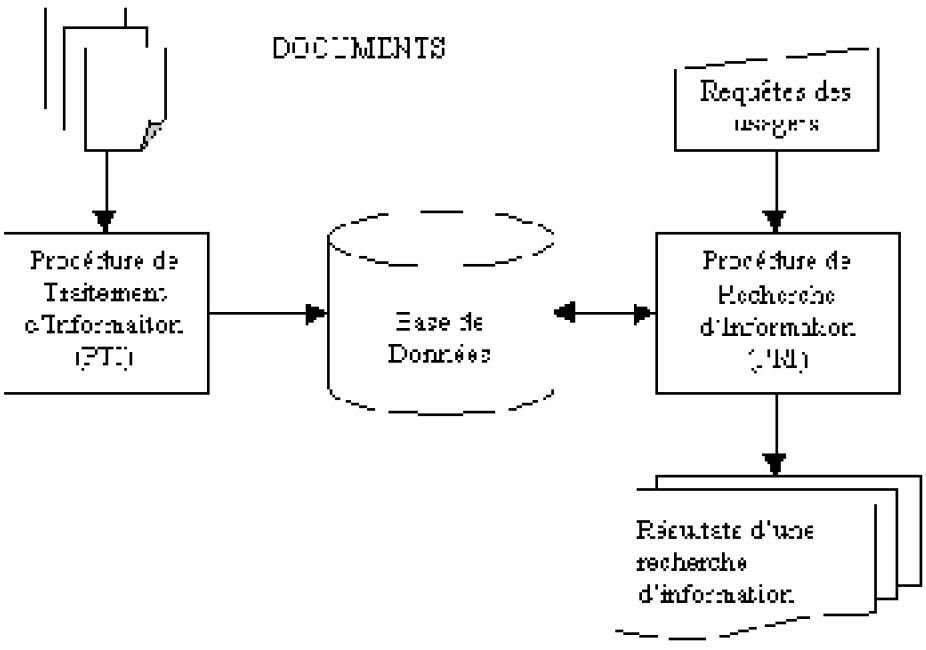
dans le chapitre 11. Ce modèle très important pour la construction de l'analyseur morpho-syntaxique, le module responsable par la reconnaissance et l'extraction de syntagmes nominaux. Il est donc prioritaire qu'on établisse d'abord le modèle maintenant et puis construire le système.

Une fois élaboré le modèle d'identification et d'extraction de syntagmes nominaux dans des textes en langue portugaise, nous sommes prêts à construire le système de recherche d'information proposé. Ainsi nous allons dédier ce chapitre à proposer, grosso modo, une structure pour ce système.

Le développement du SRI proposé sera réalisé postérieurement à cette thèse. Pourtant, nous allons montrer, en grandes lignes, comment ce système devrait être structuré.

1 Structure générale du système

Nous avons montré dans le premier chapitre, à la figure 1.2, schéma fonctionnel d'un SRI qu'il y a deux procédures distinctes. Ce schéma présente grosso modo le fonctionnement d'un SRI en montrant la composition logique et physique d'un SRI. Nous allons montrer, maintenant de manière plus synthétique, ce même schéma dans la figure 8.1.



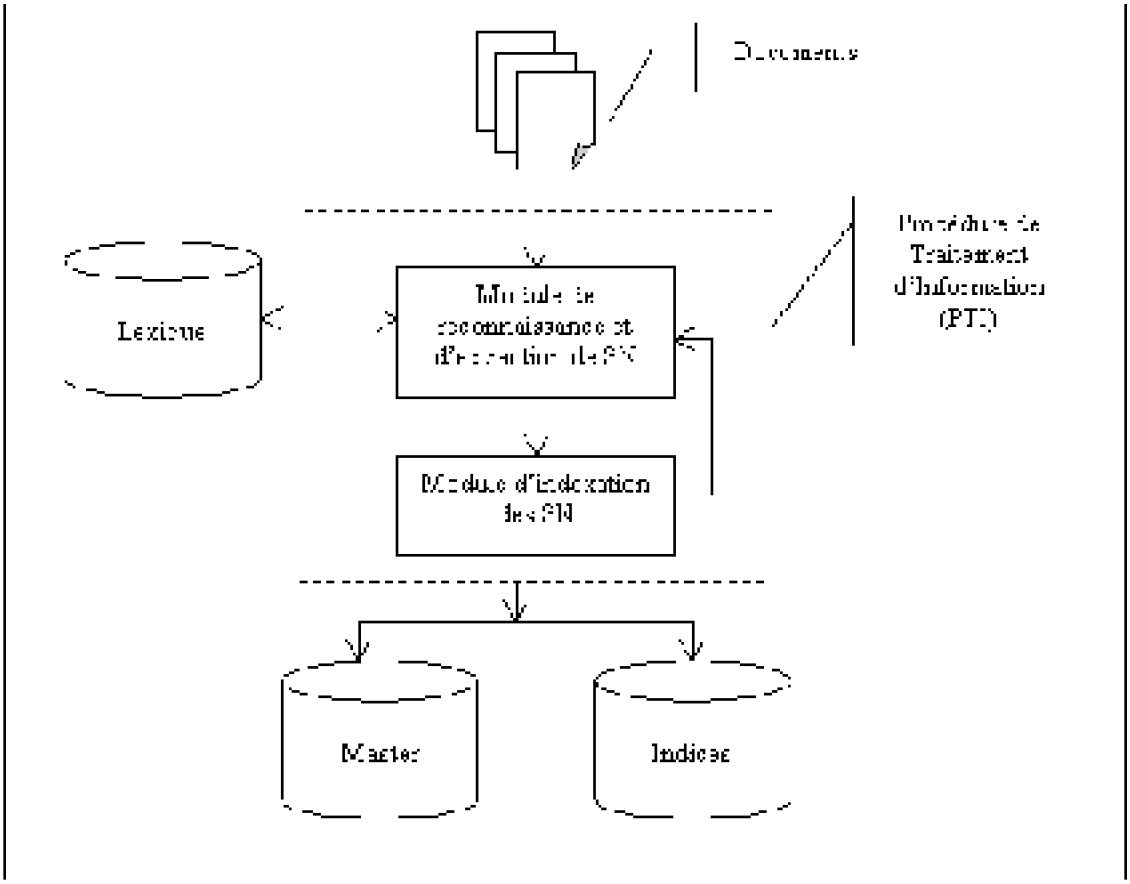
Dans la figure 8.1 on voit, de manière résumée, les procédures qui font partie d'un SRI et aussi bien les entrées et sorties (documents, requêtes et résultats d'une recherche d'information). Au milieu, on voit la base de données qui est créée et mise à jour par la procédure de traitement d'information (PTI). La base de données est, en fait, constitué de plusieurs fichiers : ce qui contient les documents de la base, les fichiers indices et les fichiers auxiliaires. Les deux procédures sont exécutées dans des moments différents, l'une lors de la préparation des documents pour leur inclusion dans la base de données,

appelée Procédure de Traitement de l'Information. C'est dans cette procédure qu'on fait l'identification et l'extraction des syntagmes nominaux (SN) des documents de façon à permettre aux usagers retrouver les documents existant dans la base de données. C'est aussi dans cette procédure qu'on organise ces SN dans une structure optimisée pour permettre aux usagers la navigation dans une base de données. C'est qu'on appelle l'indexation des documents de la base de données.

L'autre procédure est exécutée lors de la recherche d'information. On trouve dans cette procédure l'interface de recherche d'information et les procédures de parcours dans la structure de SN. Nous allons, donc, faire une indication de comment développer ces deux procédures.

1.2 Procédure de Traitement de l'Information

Cette procédure a comme tâche l'extraction des SN et leur organisation dans une structure de façon à faciliter et rendre plus rapide la recherche d'information.



Le choix des syntagmes nominaux comme moyen d'accès à l'information oblige que la Procédure de Traitement de l'Information contienne un analyseur morpho-syntaxique. C'est la manière la plus indiquée pour la reconnaissance des syntagmes nominaux car il faut reconnaître les unités lexicales qui leur appartiennent. Chaque unité lexicale a ses caractéristiques spécifiques et peut se combiner avec d'autres unités de manière à constituer un syntagme nominal. Le modèle que nous allons montrer dans les chapitres 10 et 11 est la base de cet analyseur.

Dans la figure 8.2 nous montrons la composition de la Procédure de Traitement d'Information (PTI). Cette procédure est composée, donc, par deux grands modules : 1) Module de Reconnaissance et d'Extraction de SN ; et 2) Module d'Indexation des SN.

Le Module de Reconnaissance et d'Extraction de SN est la procédure responsable par la lecture des documents, par la segmentation des textes, par la normalisation de textes et finalement par l'analyse morpho-syntaxique. Ce module utilise des informations de la base de données LEXIQUE pour la reconnaissance des unités lexicales. La flèche a été mise dans les deux sens parce que ce module peut non seulement prendre les informations des unités lexicales dans la base, mais elle peut aussi faire la mise à jour de nouvelles unités lexicales qui ne se trouvent pas encore dans la base. En ce cas, il faut les ajouter à la base, en ajoutant aussi leurs informations (variables / règles de contraints). La normalisation de texte correspond au pré-traitement des unités lexicales. En effet, autant le pré-traitement des unités lexicales que la base de données LEXIQUE seront discutés, de manière plus détaillée, dans le prochain chapitre, lors de la présentation du modèle.

La partie plus importante de ce module est l'analyseur morpho-syntaxique lequel sera le responsable effectivement pour la reconnaissance et l'extraction des SN d'un document. Pourtant, nous n'allons pas le détailler dans le cours de cette thèse. La prise de cette décision est justifiée à cause de l'avance de la technologie de l'information et de la communication. Le scénario technologique change très vite et le champ du traitement automatique de langue naturelle suit ce développement. On risque, donc, de se tromper ou de proposer une solution inadéquate à la technologie du moment du développement du SRI. Pourtant, il faut faire quelques considérations pour donner des repères à celui qui va développer ce système. Ces repères peuvent aussi être importants dans le processus de choix de la technologie qui sera utilisée dans le développement du SRI.

A suivre nous présentons ces considérations :

- Utiliser une base de données contenant les unités lexicales et leurs informations (base de données LEXIQUE). La décision d'utilisation de cette approche est justifiée dans le prochain chapitre ;
- Le module de reconnaissance et d'extraction de SN doit être capable de reconnaître d'abord le SN maximal. C'est-à-dire, le SN plus grand, ce qui ne peut pas appartenir à d'autres SN. A partir de ce SN, l'analyseur doit être capable aussi de reconnaître tous les autres SN qui y sont présents ;
- Le module de reconnaissance et d'extraction doit être capable de reconnaître les SN avec double rection et les SN qui font partie de chaque complément.

- Après la reconnaissance du SN maximal et la reconnaissance de tous les SN qui y sont présents, ce module doit les passer dans la même suite que dans le texte, avec une signalisation indiquant le début et la fin de chaque SN. On peut utiliser les parenthèses pour faire cette signalisation. Ainsi, le module d'indexation sera capable de prendre chaque syntagme nominal et aussi bien de déterminer son niveau respectif.

La module d'indexation automatique est la partie responsable pour créer, mettre à jour et organiser les syntagmes nominaux dans une structure qui puisse permettre aux usagers de trouver l'information désirée. Cette structure sera composée par les fichiers INDICES. C'est-à-dire, la structure de données qui soutiendra les syntagmes nominaux est composée non seulement par un fichier, mais par un ensemble de fichiers.

Ce module est activé par le module de reconnaissance et d'extraction de SN par le biais du passage d'une chaîne de caractères ayant la séquence de textes contenant le(s) syntagme(s) nominal(aux). Selon une des considérations qu'on a faites plus haut, cette séquence de textes a le format suivant :

(x)

Où x est un syntagme nominal et peut être composé par la forme suivante :

$x = y\ (x_1) \mid y\ (x_1)\ y\ (x_2)$

Où y est un ensemble de mots, x_1 et x_2 sont des syntagmes nominaux. Le symbole « | » indique une forme alternative. C'est-à-dire, un syntagme nominal « x » peut être composé soit par un autre syntagme nominal « x_1 » ou soit par deux autres syntagmes nominaux disposé séquentiellement comme dans un syntagme nominal avec double rection. Ce sont des formes récursives. C'est-à-dire, les syntagmes nominaux peuvent aussi être composés par des syntagmes nominaux emboîtés. Ensuite nos présentons quelques exemples pour une meilleure visualisation de cette proposition :

(L'analyse d'information)	1.
(l'étude de (l'analyse d'information))	2.
(l'acceptation de (l'information stratégique) dans (la définition de (l'avenir de (l'entreprise))))	3.

Dans l'exemple 2 on voit un exemple de syntagmes nominaux emboîtés, c'est-à-dire un syntagme nominal dans un autre syntagme nominal. Et, dans l'exemple 3 on voit le cas de syntagme nominal avec une double rection.

Un autre aspect à prendre en compte est la détermination du niveau de chaque syntagme nominal. Nous avons déjà établi cela dans la maquette construite dans le cours de cette thèse. L'énumération de chaque syntagme nominal est attribuée en ordre décroissant du syntagme maximal aux syntagmes plus petits qui sont dans ce syntagme maximal. Nous allons suivre, donc, la même approche adoptée dans la maquette. Pour exemplifier, nous allons prendre l'exemple 3, étant donné son aspect particulier d'avoir une double rection.

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, selon ce qui nous avons déjà défini dans la maquette, nous proposons que le syntagme nominal, en ce cas, appartienne à deux niveaux distincts à cause de ses rections. Ainsi, nous avons le calcul suivant :

(l'acceptation de (l'information stratégique) dans (la définition de (l'avenir de (l'entreprise))))

Ce syntagme nominal est composé de deux branches de syntagmes :

- (l'information stratégique) - Niveau 1
- (la définition de (l'avenir de (l'entreprise))) - Niveau 3
  - (l'avenir de (l'entreprise)) - Niveau 2
    - (l'entreprise) - Niveau 1

Ainsi, le syntagme plus grand :

« (l'acceptation de (l'information stratégique) dans (la définition de (l'avenir de (l'entreprise))) »

Il doit être définit comme étant un syntagme de deuxième niveau par rapport au syntagme « (l'information stratégique) » et de quatrième niveau par rapport au syntagme « (la définition de (l'avenir de (l'entreprise))) ».

La procédure d'indexation doit, donc, en utilisant les marques qui délimitent les syntagmes nominaux présents dans la chaîne repassée par le module de reconnaissance et d'extraction de SN, déterminer le niveau de chaque syntagme nominal extrait.

L'extraction de chaque SN et la détermination du niveau de chaque SN n'est pas trop difficile. Il suffit d'extraire chaque SN interne, en vérifiant l'ouverture et la correspondante fermeture de parenthèses. Chaque ensemble d'ouverture et fermeture de parenthèses détermine la présence d'un SN. On sait d'abord que la chaîne de caractères repassée par le module de reconnaissance et d'extraction de SN est déjà un SN maximal, c'est le plus grand.

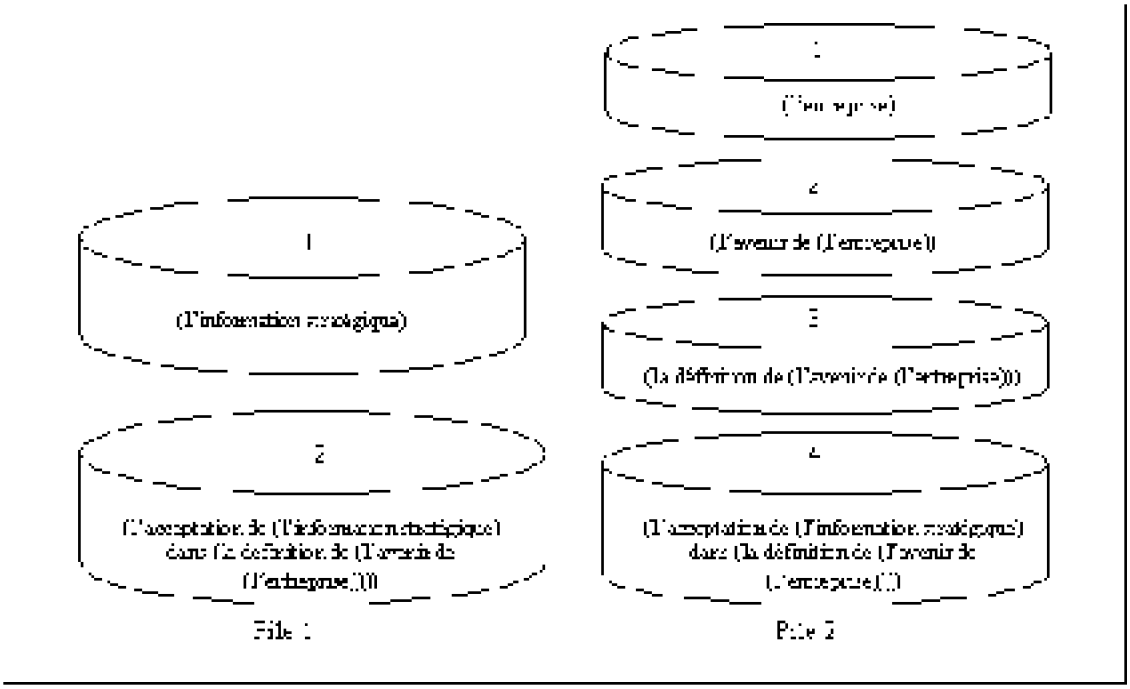
L'extraction de chaque SN et la détermination du niveau de chaque SN doit être faite en même temps. Au fur et à mesure qu'on extrait chaque SN, on les met dans une pile, aussi appelée dans la discipline de structure de données LIFO (Last In First Out), c'est-à-dire, dernier-entré-premier-sorti. Selon Christine FROIDEVAUX et al., ***« une bonne image pour se représenter une pile est une... pile d'assiettes : c'est en haut de la pile qu'il faut prendre ou mettre une assiette ! »*** [68] . Ainsi, dans notre cas, l'assiette est représentée par le syntagme nominal.

Dans le cas de SN avec double rection il faut créer le nombre de piles égal au

[68] Christine FROIDEVAUX, Marie-Claude GAUDEL et Michèle SORIA. *Types de Données et algorithmes.* PARIS : Ediscience International, 1993. p.74.

nombre de rections. Dans notre exemple, nous aurons deux piles parce que nous avons deux compléments, comme nous pouvons dans la figure 8.3.

Le schéma de la figure 8.3 montre une structure intermédiaire qui doit être utilisée lors de l'extraction de chaque SN et la détermination de leur niveau. Le même schéma montre aussi que le niveau est donné de manière croissante du SN plus petit ou le plus interne au SN le plus grand, soit le SN maximal. Ce même schéma montre aussi que dans un SN avec double rection, il y aura deux piles car dans ce SN il y a deux rections. Cela nous montre que, en ce cas, le SN maximal est déterminé comme un SN de deuxième niveau par rapport au SN plus petit, « l'information stratégique ». Et, le même SN maximal est du quatrième niveau par rapport au SN plus petit, de la deuxième rection, « l'entreprise ». Cela montre que, dans la structure de données qui va stocker les indices, le SN maximal, en cas de double rection, doit apparaître autant de fois que le nombre de rections. Là, il doit apparaître autant dans la structure de rapport entre les SN de deuxième niveau que dans celle de quatrième niveau.

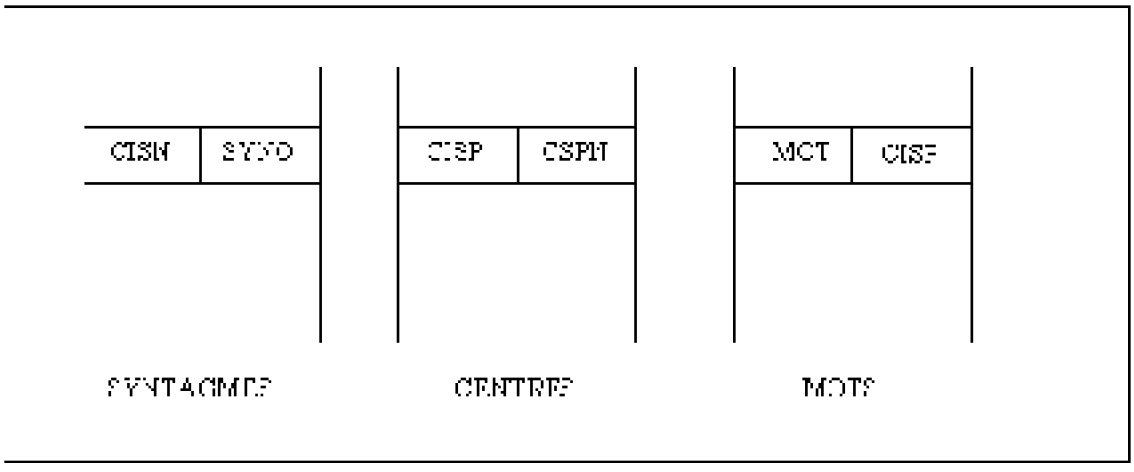


Les algorithmes pour créer, empiler et dépiler une pile sont déjà bien connus dans le milieu informatique, et ils sont aussi très simples à implémenter. Nous n'allons pas les expliciter ici. On peut les trouver dans l'ouvrage de Christine FROIDAUX et al., « Types de Données et Algorithmes », aux pages 74-76.

Le prochain pas est la construction des structures de données pour stocker les syntagmes nominaux, ceux qui vont supporter la navigation pour la recherche d'information. Tout d'abord, nous avons confirmé la faisabilité de développer un Système de Recherche d'Information assistée par Ordinateur basé sur un logiciel de gestion de bases de données relationnelles. Les tableaux avec les rapports entre les plusieurs niveaux de syntagmes nominaux ont été montrés dans le chapitre 4 - Développement de la maquette, de cette thèse.

La maquette a été développée en utilisant le logiciel Access de la Microsoft. Ce logiciel a dans son corps des routines de mise à jour des tables et des fichiers. Il n'y a pas de problèmes de mise à jour. Pourtant, ce logiciel est trop limité pour faire un système professionnel de recherche d'information. Il vaut mieux construire un système en utilisant des langages de haut niveau comme le C++ ou Java. Pour cela il faut construire toutes les structures de données à l'aide d'une méthode d'accès comme l'ISAM (Indexed Sequential Access Methode) ou l'arbre binaire ou d'autres méthodes performantes existantes dans le marché. Cette méthode permettra de faire l'accès à l'information directement, sans faire une lecture séquentielle à chaque syntagme nominal. Normalement on trouve ces méthodes dans les logiciels de gestion de fichiers. Nous n'allons pas faire du souci au problème d'accès en disque à l'information parce que nous pouvons trouver dans le marché de bons logiciels de gestion de fichiers. Ce sont des outils de développement. Ils sont déjà bien connus. Par contre, il faut dessiner les structures de données, comment on fait les liaisons entre chaque structure, et comment se fait le rapport entre les SN d'un niveau donné avec d'autres SN d'un niveau plus haut. C'est dans ce cadre là qu'on doit travailler maintenant.

Une fois connus les syntagmes nominaux et leur niveau, il faut déterminer le centre du syntagme nominal de premier niveau. La détermination du centre de syntagme nominal de premier niveau est importante en ce moment là parce que la recherche d'information, dans le modèle proposé, commence par ce centre. Les indications de comment déterminer ce centre seront données dans le chapitre 10, dans la section 3.3.7. Dans cette section, nous discutons le problème de détermination du centre de syntagme nominal du premier niveau lorsque avons un syntagme composé par une expansion prépositionnelle.



L'ensemble de structures de données nécessaires pour stocker et organiser les syntagmes nominaux a comme base trois fichiers principaux (voir la figure 8.4), définis comme étant :

- SYNTAGMES - Ce fichier contient tous les syntagmes nominaux extraits des documents qui appartiennent à la base de données. Il contient les champs suivants : le code d'identification de syntagme nominal (CISN), le syntagme nominal. La clé primaire est le syntagme nominal (SYNO) ;

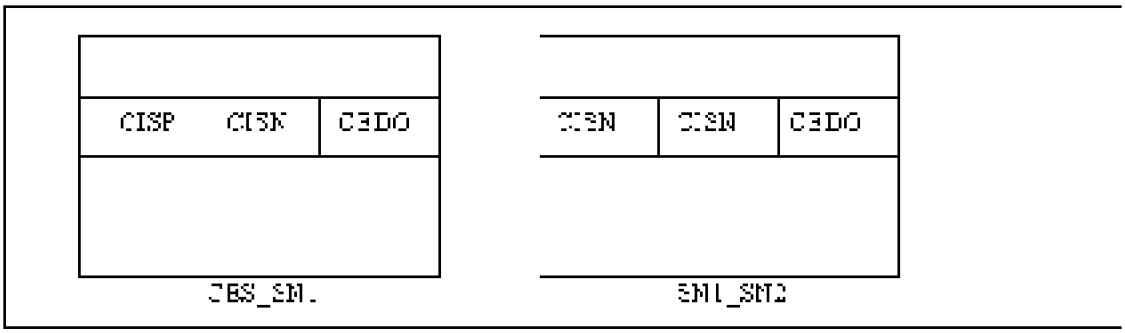
- CENTRES - Ce fichier contient les centres de syntagme nominal de premier niveau. Ce fichier contient les champs suivants : le code d'identification de centre de syntagme nominal de premier niveau (CISP), le centre de syntagme nominal de premier niveau (CSPN). La clé primaire est le champ CISP ;
- MOTS - Ce fichier contient les mots équivalents ou synonymes des centres de syntagme nominal de premier niveau. Il contient les champs suivants : le mot, le CISP. La clé primaire est le mot.

Les autres structures de données sont générées pour soutenir la recherche d'information. Ils sont donc créés à partir des données stockées dans les fichiers SYNTAGMES et CENTRES et aussi à partir des données présentes dans les piles. Les piles donnent des informations sur le rapport entre les syntagmes nominaux présents dans chaque pile.

La recherche commence par le centre de syntagme nominal fourni. En effet, l'usager peut fournir un mot qui ne se trouve pas parmi les centres de syntagme nominal, mais peut être équivalent ou synonyme. C'est pour cela que nous avons défini un fichier appelé MOTS. Ce fichier comme nous avons déjà dit contient des mots qui sont équivalents ou synonymes d'un centre de syntagme nominal. C'est pour cela que la recherche, en effet commence par trouver le CISP (code d'identification d'un centre de syntagme nominal) en faisant accès au mot dans le fichier MOTS.

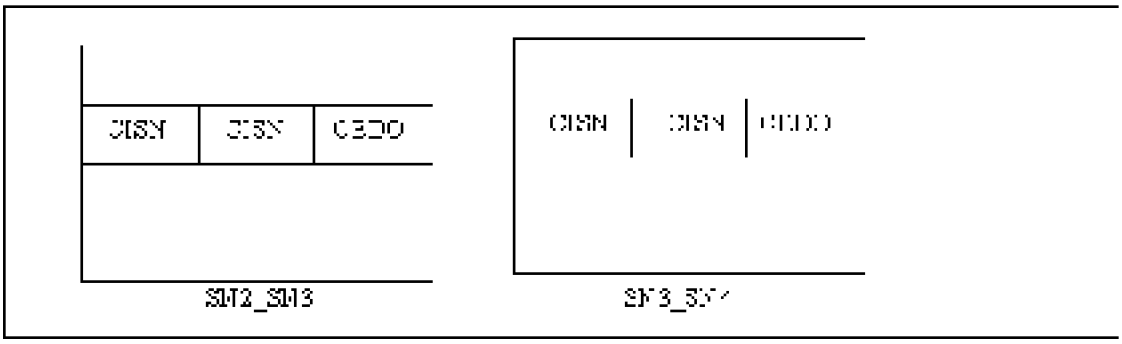
A partir de ce CISP, le système doit faire un accès à une structure qui fait le rapport entre les centres de syntagme nominal de premier niveau et les SN de premier niveau. C'est pour cela qu'il faut avoir un fichier ayant le CISP et le CISN comme clé primaire. Ainsi, nous avons dessiné le fichier appelé CES_SN1 - Rapport entre les centres de syntagme nominal et les syntagmes nominaux de premier niveau. Ce fichier contient les champs suivants : CISP - code d'identification de centre de syntagme nominal de premier niveau, CISN - code d'identification de syntagme nominal, CEDO - code de l'ensemble de documents d'où on a extrait un syntagme nominal donné.

Dans la figure 8.5 nous montrons la spécification des fichiers CES_SN1 et SN1_SN2. Le fichier SN1_SN2 contient le rapport entre les syntagmes nominaux de premier niveau avec les syntagmes nominaux du deuxième niveau. Ce fichier permettra au système de retrouver, à partir d'un SN de premier donné, les SN de deuxième niveau. Les deux fichiers sont pareils ainsi que les autres fichiers de rapport entre les syntagmes nominaux.

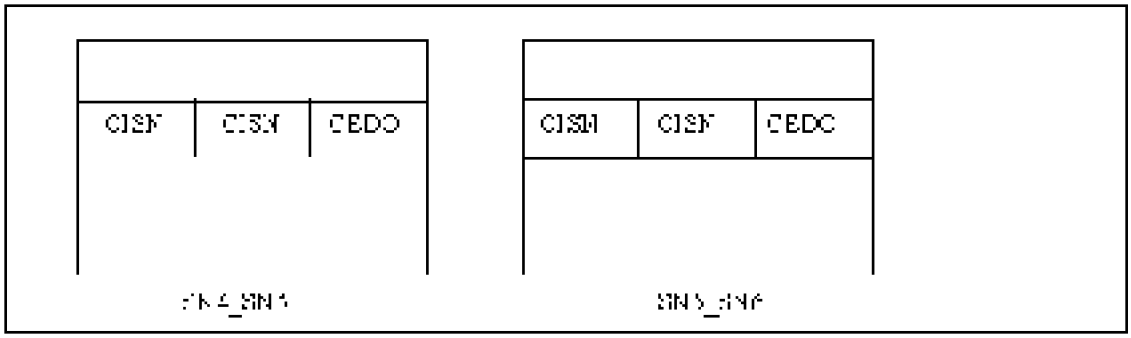


Dans les figures 8.6, nous montrons respectivement le fichier avec le rapport entre les SN de deuxième niveau et ceux de troisième niveau et le fichier avec le rapport entre

les SN de troisième niveau et ceux de quatrième niveau. De même, dans la figure 8.7 nous montrons le fichier qui contient le rapport entre les SN de quatrième niveau et ceux de cinquième niveau. Et, finalement, nous montrons aussi le fichier qui contient le rapport entre les SN de cinquième niveau et ceux de sixième niveau. Comme nous pouvons constater, les spécifications sont pareilles, il n'y a aucune différence, sauf les niveaux des SN, mais les codes sont toujours les mêmes, ceux qui identifient les syntagmes nominaux.

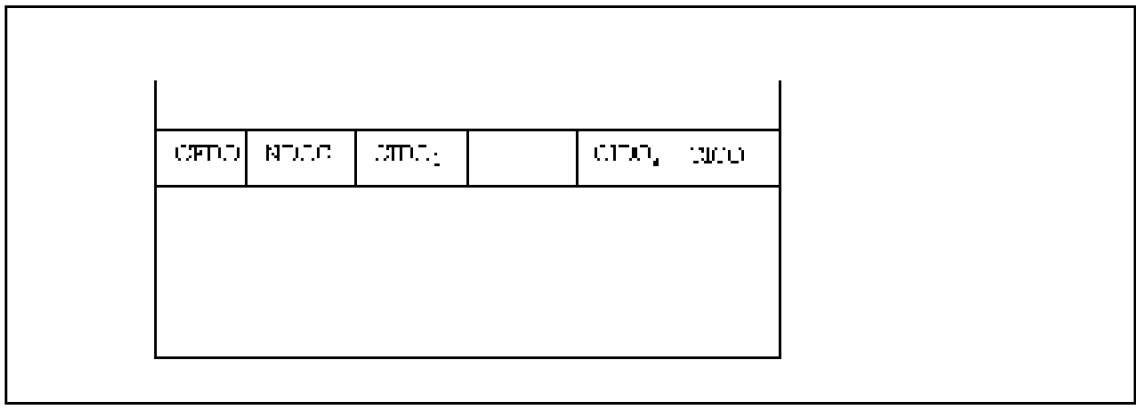


De la même façon que le fichier SN1_SN2 permet au système retrouver tous les SN de deuxième niveau, à partir d'un SN de premier niveau donné, le fichier SN2_SN3 permet au système de retrouver tous les SN de troisième niveau, à partir d'un SN de deuxième niveau donné. Cette idée est valide pour tous les fichiers de rapport entre les SN d'un niveau donné et les SN de niveau supérieur consécutif.



Dans tous les fichiers la mise à jour est faite à la fin du fichier. C'est-à-dire, les inclusions de nouveaux codes d'identification de syntagmes nominaux sont faites à la fin.

Le code d'identification de centre de syntagme nominal de premier niveau (CISP) et le code d'identification de syntagme nominal (CISN) sont de numéros séquentiels. Ainsi, les inclusions de nouveaux syntagmes nominaux et de nouveaux centres de syntagmes nominaux de premier niveau sont aussi faites à la fin de leur fichier respectif.

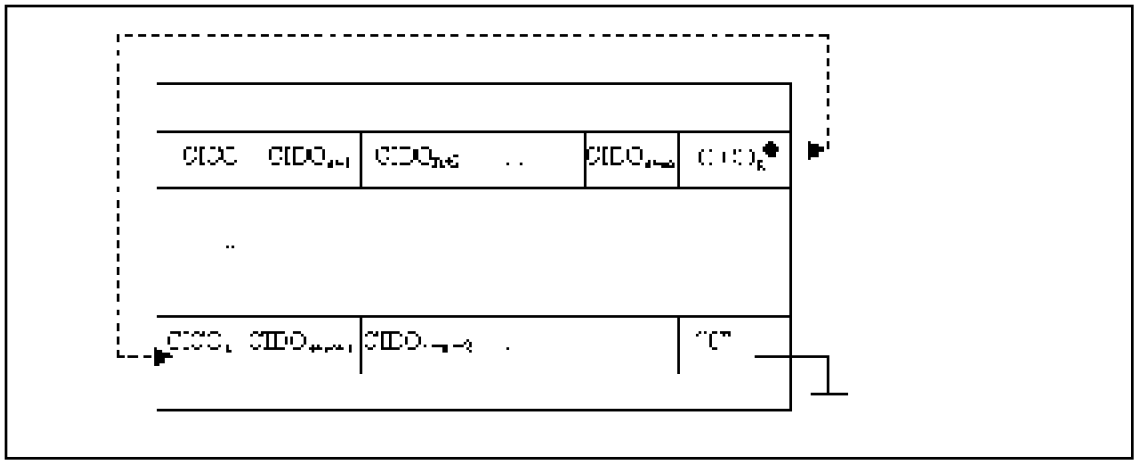


De plus, il faut avoir un fichier où seront stockés les numéros de chaque document d'où les syntagmes nominaux ont été extraits, appelé ENS_DOC. Ce fichier permettra au système de retrouver les documents à partir d'un SN donné. C'est à dire, à partir d'un SN le système pourra rencontrer les documents d'où ce SN a été extrait. Ce fichier contient les champs suivants : CEDO - le code d'identification d'ensemble de documents, NDOC - le nombre de documents d'où on a extrait un SN donné, CIDO - code d'identification de document. Le champ CIDO est répété autant de fois que le nombre de documents d'où on a extrait un SN donné. La figure 8.8 montre la spécification de ce fichier.

Le code CEDO est un code séquentiel et il est aussi la clé primaire de ce fichier. La mise à jour se fait toujours à la fin du fichier, et l'inclusion d'un nouveau document pour un syntagme nominal donné est faite toujours à la fin de chaque registre.

Il est claire qu'on ne pourra pas inclure indéfiniment des nouveaux CIDO dans un registre du fichier ENS_DOC. C'est pour cela que nous avons prévu une insertion de « n » CIDO par registre. C'est-à-dire, on peut enregistrer, au maximum, « n » documents contenant un SN donné, dans chaque registre. La valeur de « n » dépend de la taille du registre de ce fichier. Nous ne pouvons pas établi maintenant la valeur pour cette taille. C'est un détail de projet de fichier et dépend des caractéristiques du gestionnaire de fichiers choisi. Si on trouve plus de n documents contenant un SN donné, il faut remplir le champ CICO pour indiquer la continuation de ce registre dans le fichier ENS_DOC_BIS. Le CICO, code d'identification de continuation d'enregistrement de CIDO, est un code séquentiel et aussi la clé primaire de ce fichier.

Le fichier ENS_DOC_BIS a comme but stocker les CIDO qui ne pourraient pas être enregistrés dans le fichier ENS_DOC. La spécification de ce fichier est montrée dans la figure 8.9.



Le fichier ENS_DOC_BIS est composé de registres contenant les champs : CICO - code d'identification de continuation d'enregistrement de CIDO, CIDO et d'un champ de liaison pour la continuation éventuelle d'enregistrement de CIDO, dans un nouveau registre dans le même fichier. Le code CIDO peut être répété « m » fois dont la valeur dépend de la taille du registre. Si la quantité de CIDO à être enregistrée est plus grand que « m », le système doit alors créer un autre registre dans le même fichier pour continuer d'enregistrer les CIDO, comme on a montré dans la figure 8.9. Le champ CICOK — ce qui fait la liaison à d'autres registres, ce qui se trouve à la fin du registre — du dernier registre de cette séquence de CIDO a comme valeur zéro (0). Cette valeur signale qu'il ne fait aucune liaison à d'autres registres. Elle indique aussi que ce registre est le dernier pour un de l'ensemble de documents d'où on a extrait un SN donné. La création de ce fichier résout les possibles contraintes dues à la quantité de documents qui puissent avoir un SN donné.

Nous allons présenter maintenant une autre structure de données laquelle a une fonction auxiliaire. Elle n'est pas liée directement à la recherche d'information, mais elle est très important pour aider les usagers à visualiser un SN donné dans les documents trouvés. C'est-à-dire, il est intéressant aux usagers de voir, dans les documents trouvés à partir d'une recherche d'information, les endroits d'où a été extrait le SN qu'il a utilisé pour retrouver ce document. Ces informations peuvent être utiles à l'usager pour savoir le(s) endroit(s), dans un document retrouvé dans une recherche, où se concentre le SN utilisé dans la recherche. Une autre application de ces informations est dans les champs de la recherche d'information. Les parties où se concentra un SN donné peut être utiles pour l'établissement de l'unité d'un document. C'est-à-dire, l'usage de ces informations peut nous amener à prendre en compte comme un registre d'information, non le document total (livre, article etc.), mais partie(s) de ce document dont l'occurrence d'un SN donné est plus concentrée. Une autre application d'utilisation de ces informations est la détermination d'ordre de pertinence d'un document par rapport au pourcentage d'occurrence d'un SN donné. C'est une chose pareille à ce que les modèles traditionnels d'indexation font pour les mots. Le calcul d'ordre de pertinence en utilisant l'occurrence de mots dans un document.

Ainsi, nous proposons que le module de reconnaissance et d'extraction de SN fasse le repérage du numéro de la ligne, du document, d'où un SN donné a été extrait et le

repasse au module d'indexation. Ce module à son tour l'enregistre dans le fichier POS_SN. Ce fichier aura comme clé primaire le code CISN - code d'identification de Syntagme nominal. De plus, ce fichier ne gardera que des numéros de lignes (NL) d'où le respectif SN a été extrait. Etant qu'il aura des SN qui auront beaucoup d'occurrence dans un document, il est nécessaire la création d'un fichier de continuation d'enregistrement de ces adresses (numéros de ligne d'où le SN a été extrait), NL. La structure de donnée nécessaire pour supporter ces informations est pareille à celle présenté dans les figures 8.8 et 8.9.

Pour une question de clarté nous montrons dans les figures 8.10 et 8.11 les fichiers qui font partie de la structure de données dessinées pour gardes les adresses des SN de la base de données. Dans ces fichiers, CINL indique la liaison d'un registre au registre de continuation d'enregistrement de numéros de lignes. CINL est aussi la clé primaire du fichier POS_SN_BIS.

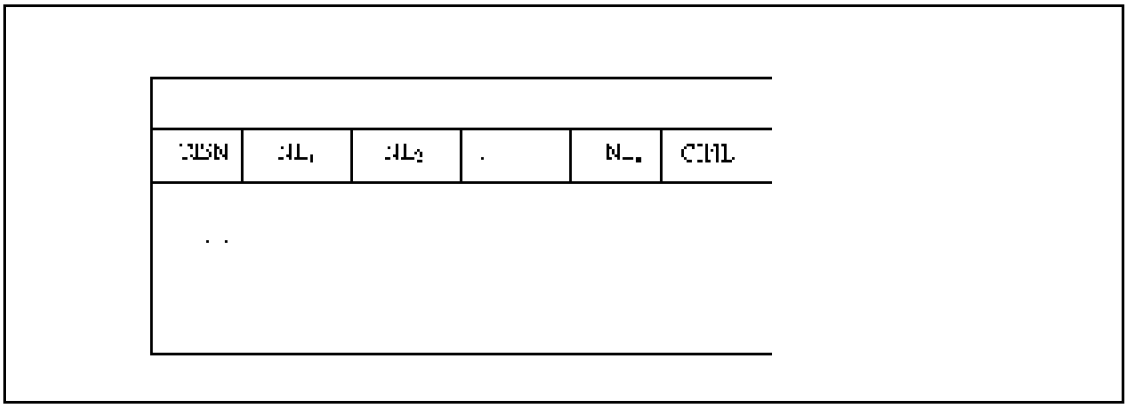


FIG08-11.gif

En ce moment, nous n'allons pas faire du souci aux problèmes de tailles des divers codes et ni des divers registres de chaque fichier. Ce problème doit être réglé lors de l'implémentation car il faut connaître le système opérationnel, même le gestionnaire de fichiers et bien aussi la plate-forme de hardware pour laquelle sera développé ce nouveau SRI. Ce n'est pas le moment, donc, de calculer les paramètres relatifs aux tailles des fichiers et aussi de champs.

Il manque, maintenant le fichier qui contient les textes des documents dans la base de données. Ce fichier est composé par les champs : CIDO - code d'identification de document, le texte du document. Ici, nous pouvons avoir aussi, au lieu du texte du document, une référence du fichier qui le contient. Mais cela dépend du système de gestion de fichier, c'est-à-dire des limitations du gestionnaire car il est fort probable que la taille des textes de documents sera trop grand. De toute façon, il me semble que la solution de mettre une référence à un fichier qui contient le document est une solution plus générique et indépendante du système de gestion de fichiers.

1.3 Procédure de Recherche d'Information

Cette procédure est composée de deux procédures ou routines : 1) Interface de Recherche d'Information ; et 2) Le module de navigation dans la structure de syntagmes

nominaux. La figure 8.12 montre le schéma fonctionnel de la procédure de recherche d'information.

FIG08-12.gif

Selon la figure 8.12, le module Interface de Recherche d'Information est responsable pour interagir avec l'usager. C'est ce module qui reçoit la demande de recherche d'information et la présentation de toutes les réponses du SRI à une demande de l'usager. Cette interface a aussi une interaction avec le module de navigation dans les structures de SN de la base de données. Ces interactions sont faites toujours dans les deux sens.

Pour la construction de l'Interface de Recherche d'Information, il faut prendre en compte les recommandations présentées dans cette thèse à la fin du chapitre premier, dans sa conclusion.

Dans un autre côté, la suite de dialogue entre l'usager et l'interface de recherche d'information peut utiliser le même schéma montré dans le chapitre 4, dans la figure 4.1. Il est évident qu'à cause d'obtenir plus de convivialité, au moment du développement du nouveau SRI, on peut changer quelque chose du schéma présenté. Mais grosso modo le schéma à utiliser est celui de la figure 4.1, il représente le modèle de SRI qui nous venons de proposer dans cette thèse.

FIG08-13.gif

En ce qui concerne le module de navigation, il suit la demande de l'usager. S'il veut demander une recherche, l'interface de recherche d'information lui demande un mot ou un centre de syntagme nominal. A partir de là, le module cherchera à trouver tous les SN de premier niveau qui ont ce centre de syntagme. L'accès à ces syntagmes nominaux sera fait à partir du fichier CES_SN1. L'accès aux SN de niveaux consécutifs est fait à travers les fichiers correspondant à chaque rapport entre les SN. Exemple : SN1_SN2 pour accéder aux SN de deuxième niveau, SN2_SN3 pour accéder aux SN de troisième niveau et ainsi de suite. C'est-à-dire, l'accès aux SN d'un niveau est fait de manière séquentielle et consécutive, ce n'est pas possible de sauter un niveau donné de SN.

Dans la figure 8.13 nous présentons, de manière schématique, le parcours que la structure de SN dessinée permet aux utilisateurs. Les fichiers de rapport entre les divers niveaux de SN sont montrés de manière simplifiée, le champ CEDO n'apparaît pas dans le schéma. Nous ne l'avons pas spécifié dans chaque fichier car il n'est pas nécessaire à ce moment là.

2 Conclusion

La question principale, quand on propose un nouveau système, est qu'est-ce qu'il faut pour construire un système de recherche d'information assistée par ordinateur ? Nous essayons de répondre à cette question de manière à donner une idée de la structure et de touts les éléments nécessaires pour un tel système. Il est vrai que nous n'avons pas détaillé complètement le projet, mais ce n'était pas le but de ce chapitre.

Nous avons montré ici une séquence de procédures nécessaires au développement

du nouveau Système de Recherche d'Information proposé par ce travail. C'est en réalité un ensemble minimum de procédures qui amène à la construction d'un nouveau SRI. Il est évident que nous pouvons concevoir d'autres procédures pour l'amélioration du système. Par exemple, la partie relative à l'interface de recherche d'information, nous avons proposé une suite d'interaction entre l'usager et l'ordinateur. Ce dialogue permet à l'usager de naviguer dans la structure de syntagmes nominaux, mais rien ne nous empêche de construire une nouvelle interface, aussi interactive, mais avec des facilités graphiques différentes de celles que nous avons utilisées.

Une autre possibilité d'aide aux usagers dans la procédure de recherche d'information, utilisant les SN comme moyen d'accès à l'information, est la création d'un fichier contenant des SN équivalents ou synonymes à ces présents dans le fichier de SN extraits. La présentation de SN e ses équivalents ou synonymes peut aider les usagers dans une procédure de recherche d'information. Cependant, la création de ce fichier est un peu difficile car il faut l'intervention humaine pour le créer. A part cette difficulté, la création de ce fichier serait très utile pour la recherche d'information. Cela aiderait à diminuer le taux de silence.

Il faut faire une autre remarque, celle sur le dimensionnement des tailles des champs dans les fichiers et même de celles des fichiers. C'est un détail d'implémentation et qui dépend du système de gestion de fichiers qui sera utilisé et de la plate-forme de hardware. Ainsi, le but de ce chapitre a été de montrer, en grandes lignes, une approche de développement basic du nouveau SRI.

Ce chapitre peut être vu comme ce qui fait l'intégration de tous les points abordés par cette thèse. C'est l'intégration entre la partie pratique et la partie théorique. En effet, ce chapitre et les deux autres qui suivront constituent, sinon le projet logique, du moins la base pour la construction d'un système de recherche d'information assistée par ordinateur.

« La question des parties du discours se situe, le plus souvent de manière implicite, à l'articulation du lexique et de la syntaxe . » Michel LE GUERN. « Parties du discours et catégories morphologiques en analyse automatique ». Les Classes de Mots. p. 207

Chapitre 9 Grammaire de Référence

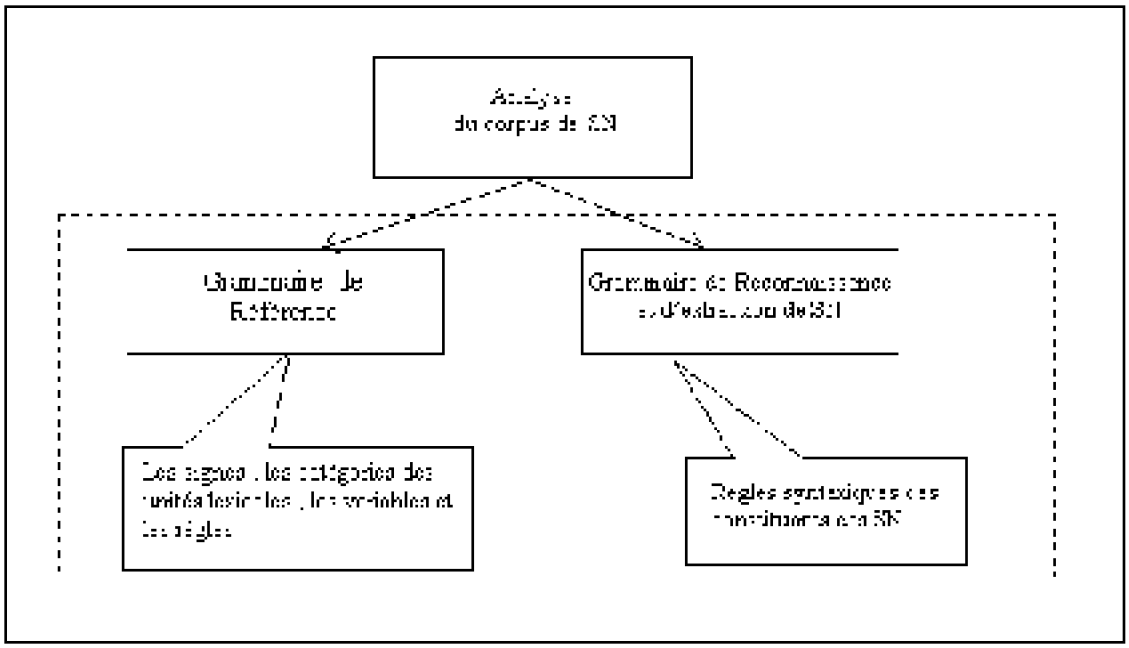
1 Considérations préliminaires

Tout d'abord, le but de cette partie de la thèse est d'établir un modèle pour la construction d'un système automatique de reconnaissance et d'extraction des syntagmes nominaux (SN) dans un ensemble de documents textuels écrit en langue portugaise. Ce modèle sera composé de deux modules : 1) module de description lexicale, appelé Grammaire de Référence ; et 2) module Grammaire de Reconnaissance et d'Extraction de SN. La figure

9.1 montre le schéma de développement de ce modèle.

La démarche utilisée pour la construction de ce modèle est partie de l'analyse du corpus des SN extraits lors de la construction de la maquette développée dans le cadre du cours de DEA [69] .

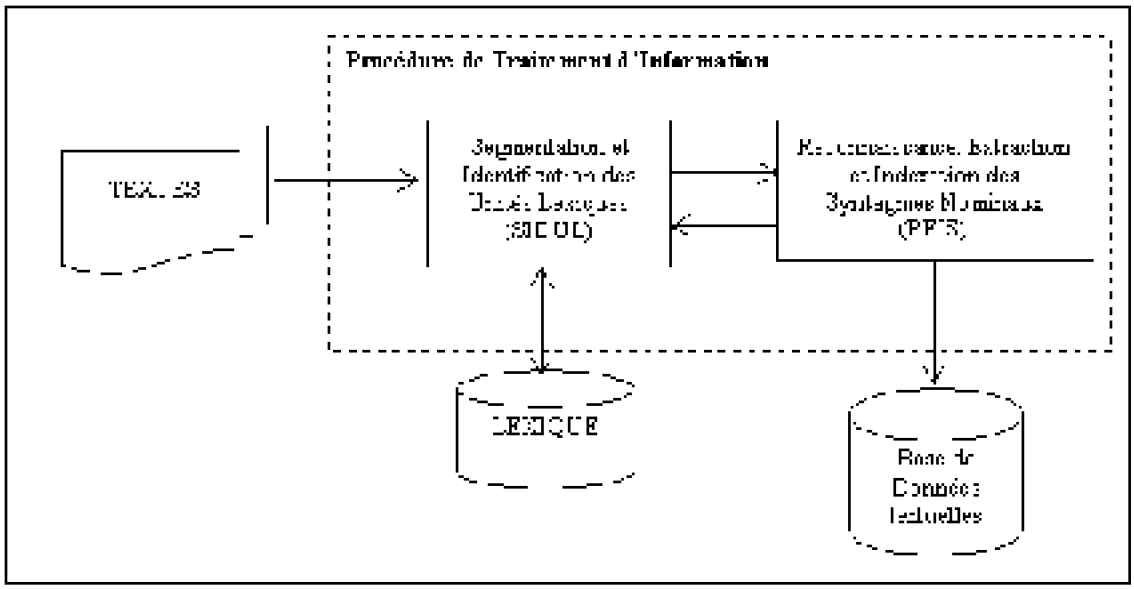
La Grammaire de Référence contient les caractéristiques, des unités lexicales, nécessaires à la reconnaissance d'un SN. Comme caractéristiques, grosso modo, on trouve : 1) la catégorie de l'unité lexicale ; 2) variables et règles nécessaires pour lever l'ambiguïté sur une unité lexicale au moment de la reconnaissance des unités lexicales dans la procédure d'extraction des SN. La Grammaire d'extraction contient les règles syntaxiques pour la reconnaissance et l'extraction des SN.



Ce modèle servira de base pour le développement de l'analyseur morpho-syntaxique. La figure 9.2 montre, grosso modo, le schéma de fonctionnement de la Procédure de Traitement d'Information (PTI). Il y aura deux grands modules : 1) Module de Segmentation et d'Identification des Unités Lexicales (SIDUL) ; 2) Module de Reconnaissance, d'Extraction et d'Indexation des Syntagmes Nominaux (REIS). Le module SIDUL segmente le texte, utilise et fait la mise à jour de la base de données LEXIQUE pour l'identification des unités lexicales. Le module REIS utilise les informations de l'unité lexicale trouvées par le SIDUL pour la reconnaissance et extraction des syntagmes nominaux. Le module REIS est, donc, composé de deux sous-modules, un qui est responsable pour l'analyse morpho-syntaxique et l'autre qui est responsable pour l'organisation des SN, c'est-à-dire, pour l'indexation automatique. Le schéma de la figure 9.2 est légèrement différent de celui de la figure 8.2 car dans cette section il fallait mettre en évidence la segmentation de texte et l'identification.

[69] Hélio KURAMOTO. *Maquette d'un système de recherche d'information en utilisant des syntagmes nominaux.* Villeurbanne, 1995. Mémoire du DEA. École Nationale Supérieure des Sciences de l'Information et des Bibliothèques.

La base de données LEXIQUE contiendra toutes les unités lexicales, trouvés dans les textes analysés, avec leurs caractéristiques. Celles qui sont décrites dans la Grammaire de Référence, que nous définirons dans ce chapitre. La mise à jour est faite au fur et à mesure que les textes seront traités et indexés. Le début sera un peu difficile car il y aura une grande quantité d'unités à caractériser et à inclure dans la base LEXIQUE. A partir d'une certaine quantité de documents traités, le volume d'unités lexicales à inclure dans la base LEXIQUE doit diminuer sensiblement.



La proposition de ce travail est de garder les unités lexicales dans toute leur forme propre, on ne fera pas ce qu'on appelle de lemmatisation [70] . Aujourd'hui les modules de mémoires secondaires (les disques durs) ne sont pas chers et sont très rapides. Ainsi, on peut utiliser un logiciel de gestion de bases de données relationnelles pour créer cette base de données au lieu de développer un analyseur morphologique basé sur la lemmatisation. Une autre solution serait de chercher un système qui ferait la segmentation du texte et l'analyse morphologique. Or, cela peut ne pas être une bonne solution car il faut faire beaucoup d'efforts pour s'adapter aux contraintes du logiciel choisi. Ce genre de logiciel est normalement une sorte de boîte noire. De plus, on reste toujours dépendant du constructeur du logiciel. Dans l'approche présentée ici, le module de reconnaissance et d'extraction des syntagmes nominaux fait un échange fréquent avec le module SIDUL. Ainsi, il faut que le module SIDUL (à développer) ou un logiciel analogue (déjà prêt) soit capable de s'intégrer dans le module REIS. Il ne s'agit pas de deux procédures isolées. Ainsi, la meilleure solution est de concevoir et construire le module SIDUL à partir d'une application basée sur un système de gestion de bases de données relationnelles. La

[70] Selon Georges Mounin, lemmatisation est une sorte de « *opération consistant à regrouper les formes occurrentes d'un texte ou d'une liste sous des adresses lexicales. On distingue en général deux étapes : 1) le regroupement des formes fléchis sous la forme type leur servant d'adresse lexicale ou lemmatistion à proprement dit ; 2)la séparation des formes servants d'adresses lexicales quand elles sont homographes (ex. : voile, s.m,, et voile, s.f.). Certains auteurs et praticiens utilisent aussi le terme de lemmatisation pour désigner tout regroupement (lexies complexes, paradigmes flexionnelles…). On préfère de plus en plus utiliser ici le mot indexation*. ». Georges Mounin. *Dictionnaire de la linguistique*. Paris : Quadrige / Presses Univesitaires de France, 1993. p. 200.

contrainte est que le logiciel de gestion de base de données utilisé soit compatible avec le langage de développement des deux modules : SIDUL et REIS. C'est-à-dire que le langage utilisé, pour la programmation de ces modules, puisse donner accès aux données créées par le système de gestion de bases de données relationnelles.

Le développement d'un système complet de segmentation et d'identification d'unités lexicales, en utilisant le principe de lemmatisation, ne fait pas partie de cette thèse car ce développement demanderait un effort équivalent au développement d'une autre thèse. Cependant, la solution d'utiliser un logiciel de gestion de bases de données relationnelles nous semble bonne. Ce qui importe dans cette recherche c'est la définition d'un analyseur morpho-syntaxique capable de fournir des informations pour l'identification et l'extraction des syntagmes nominaux.

La Grammaire de Référence des unités lexicales est important pour la création de la base de données LEXIQUE, tandis que les règles de la grammaire d'extraction de syntagmes nominaux feront partie des algorithmes du REIS, le module qu'inclut l'analyseur morpho-syntaxique.

Dans ce chapitre nous présenterons la grammaire de référence des unités lexicales et nous proposerons une structure de données pour la construction de la base de données LEXIQUE.

2 Définitions préliminaires

Le but principal de ce travail est de reconnaître et d'extraire les syntagmes nominaux d'un texte en langue portugaise. Ce texte doit être dans un format permettant de le lire et de l'analyser. C'est pourquoi, nous pensons qu'il doit être en format libre, on définit que le texte est en format texte libre utilisant le code ASCII [71] . Cependant, pour un système de recherche d'information il nous semble nécessaire d'utiliser un format qui puisse expliciter l'organisation logique d'un document. Ainsi, nous proposons l'utilisation d'un format compatible avec le format SGML.

Standard Generalized Markup Language (SGML) signifie, en Français : langage normalisé de balisage généralisé. Selon Eric VAN HERWIJNEN [72] , ***« cette norme permet l'échange de documents et est destinée plus particulièrement au domaine de l'édition mais peut aussi être appliquée au domaine bureautique et à l'industrie. Les documents SGML ont une structure décrite rigoureusement, qui peut être analysée par ordinateur et être facilement comprise par un être humain. »*** .

Selon Victor SANDOVAL [73] , ***« SGML est un langage pour écrire des applications spécialisées. Le principal objectif de SGML est de définir des structures logiques, mais il permet aussi de définir d'autres structures telles que les structures***

[71] ASCII code de caractères utilisés par les ordinateurs PC. Il est la sigle de American Standard Code for Information Interchange.

[72] Eric Van Herwijnen. *SGML Pratique*.Paris : International Thomson Publishing France., 1995. P. 3.

[73] Victor Sandoval. *SGML : un outil pour la gestion électronique de documents.* Paris :Hermés, 1994. P. 33.

hiérarchiques de données. » .

Ainsi, SGML est un outil, aussi classé comme langage, permettant de décrire un texte d'un document. Il décrit non seulement un texte mais il le fait de manière structurée. C'est un langage orienté à balisage. Le mot balisage est utilisé pour désigner les instructions ou les caractéristiques que les anciens éditeurs écrivaient sur un texte à imprimer, par exemple des informations comme : le nom de la police de caractère, la taille des caractères, l'aspect (normal, gras, italique), la justification et l'indentation du texte et d'autres caractéristiques. L'introduction de l'ordinateur dans l'industrie de l'édition a précipité l'apparition de plusieurs langages de balisages analogues au système manuel. ***« En 1978, un groupe de travail ANSI (American National Standard Institute) (X3 J6) fut formé afin de définir un format non ambigu pour l'échange de textes et un langage de balisage, qui serait suffisamment riche pour permettre tout traitement (futur). Au début des années quatre-vingt, ce travail fut transféré à l'ISO (International Standard Organization) dans un groupe de travail qui faisait partie du SC18 (ISO/IEC-JTC1/SC18/WG8) dont le travail donnera naissance plus tard à la norme SGML. »*** [74] .

Le SGML permet de :

1. Faciliter l'échange de documents avec n'importe quelle machine ou système opérationnel, étant donné qu'il est un standard international. Il existe déjà des logiciels capables de lire et de convertir un document SGML en un autre format ;
2. Organiser logiquement un texte dans une structure en arbre, permettant les indications référentielles du document (auteur, titre du document, sujet et d'autres références) ;
3. Séparer la forme du contenu (la forme est cachée dans les feuilles de style ou dans les macros) ;
4. Apporter plus de transparence en ce qui concerne la typographie et la mise en page.

Sans approfondir sur la description du SGML, nous pouvons faire quelques remarques montrant les avantages de l'utiliser dans une implémentation d'un système de recherche d'information. Ces avantages sont les suivants :

1. Possibilité de mettre des informations référentielles du document a être traité, comme le nom de l'auteur, le titre, les titres de sections et paragraphes ;
2. Possibilité d'inclure des marques lors de l'extraction des SN, pour les expliciter après au moment de la présentation du respectif document à l'usager. Ainsi, par exemple, on peut souligner ou mettre en caractères gras tous les syntagmes nominaux extraits d'un document ; ceci facilitera aux usagers la vision des SN extraits d'un document ainsi que ceux utilisés lors de la recherche d'information ;
3. Possibilité de créer des liens entre un syntagme et un autre, ou même entre les articles d'un même auteur. Ce qui peut donner plus de souplesse à la navigation dans l'ensemble de documents au moment de leur visualisation ;

[74] Eric Van Herwijnen. *SGML Pratique*.Paris : International Thomson Publishing France., 1995. p. 24.

4. Proximité avec le format HTML, ce qui permet à une application d'être converti plus facilement d'un environnement non-WEB à celui du WEB.

De toute façon, on listera dans la bibliographie une série d'ouvrages concernant l'outil SGML et son utilisation. Ainsi, pour une question de simplicité et en suivant le but premier de cette recherche, qui est de concevoir un modèle pour la reconnaissance et pour l'extraction des syntagmes nominaux, dans des textes écrit en langue portugaise, on définit le format de texte comme étant de format libre, sans aucune codification autre que le code de chaque caractère (ASCII).

En ce qui concerne la langue portugaise, il faut remarquer que nous allons travailler plutôt avec la langue portugaise écrite et parlée au Brésil. Il y a quelques différences entre celle-ci et celle qui est écrite et parlée au Portugal.

Nous allons d'abord adopter une notation pour expliciter les éléments, les catégories, les règles et d'autres éléments du modèle. Nous utiliserons la notation appelée BNF (Backus-Naur Form ou Backus-Normal Form). Cette notation a été développée par J. W. BACKUS [75] , et utilise les symboles suivants :

- Les noms en lettres majuscules représentent le nom d'un élément non terminal ;
- Les noms et les définitions sont mis entre '<' , '>' et en lettre minuscule ;
- Définition des symboles :

::=
ce symbole signifie : est défini comme

□
ce symbole signifie : est composé de...

|
ce symbole signifie : ou

- Exemple :

1. E : := <une suite de chiffres> On lit cette notation comme étant : le symbole E est défini comme une suite de chiffres.
2. E □ <no> On lit cette notation comme étant : E est composé d'une variable <no> ;
3. <no> □ <no> <chiffre> | <chiffre> On lit cette notation comme étant : <no> est composé d'une variable <no> suivi d'un <chiffre> ou il est composé d'un <chiffre>
4. <chiffre> □ 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 On lit cette notation comme étant : <chiffre> est composé du caractère 0 ou du caractère 1 ou du caractère 2 et ainsi de suite.

[75] Cette notation a été extrait du livre de David GRIES. *Compiler Construction for Digital Computers.* New York : John Wiley & Sons, 1971. p. 19, lequel donne le crédit intellectuel au travail original de J. W. BACKUS. « The syntax and semantics of the proposed international algebraic language of the Zurich ACM-GAMM Conference. ». *Proceedings of International Conference on Information Processing.*UNESCO :1959, p. 125-132.

<lettre> □

a | b | c | d | e | f | g | h | i | j | | k | l | m | n | o | p | q | r | s |t | u | v | y | w | z | A | B | C | D | E | F | G | H | I | J | K | L |M | N | O | P | Q | R | S | T | U | V | Y | W | Z | à | á | â | ã | ê | é | í | ú | ü | ô | ó | õ | À |Á | Â | Ã | É | Ê | Í | Ó | Ô | Õ | Ú | Ü | ç | ÇLes lettres 'k', 'w' et 'y' ne sont utilisées que dans deux cas 104 Celso CUNHA et Lindley CINTRA. Nova Gramática do Português Contemporâneo. Lisboa : Edições João Sá da Costa, 1991. p. 63 :

1. Dans la transcription de noms propres étrangers et de leurs dérivés en portugais. Exemples : Franklin, frankliniano, Wagner, wagneriano, Klabin etc.
2. Dans les raccourcis et symboles internationaux. Exemple : km (quilômetro), K (potassium), yd (jardes), w (watts), kg (kilogrammes) etc.

<signes de ponctuation> □

, | ; | . | : | ? | ! | … Ces signes peuvent être utilisés pour la segmentation de phrases. Cependant, il faut remarquer que le point '.' peut aussi être utilisé dans des sigles et abréviations. Exemples : S.A. (Société Anonyme), C.V.R.D. (Companhia Vale do Rio Doce). Pourtant, dans le cas de sigles, le problème peut être supprimé car les points sauf le dernier n'y sont pas suivis d'espace. Cependant, aujourd'hui c'est rare d'écrire les sigles avec des points entre chaque lettre. En ce qui concerne les abréviations, le problème est plus difficile à résoudre. Une solution serait de les mettre dans la base LEXIQUE, en indiquant qu'il s'agit d'une abréviation.

<tiret/trait d'union> □

- Le <tiret/trait d'union> apparaissent dans deux situations, soit dans les mots composés et verbes fléchis avec un pronom (trait d'union), soit en dehors des mots où il joue un rôle de ponctuation (tiret). La différence entre les deux rôles de ce signe est :

1. lorsqu'il joue le rôle de trait d'union, il est mis entre deux mots, sans espaces, comme dans les mots composés et aussi lorsqu'on utilise les pronoms personnels (« obliquos atonos ») équivalents aux pronoms clitiques en français, après le verbe [me, te o, a, os, as, lhe, lhes (me, te, le, la, les, las, leur, leurs)] ; Exemple : guarda-chuva (parapluie), couve-flor (chou-fleur), ofereceram-me (ils m'ont offert), levaram-no (ils l'ont amené) etc.
2. lorsqu'il joue le rôle de tiret, il est entouré par deux espaces ou par un espace et une ponctuation. | □-□ | ou | □-,| où le signe □ est utilisé pour désigner un espace.

<ponctuation double> □

« » | ( ) |

<signe de délimitation> □

□ On considère l'espace (□) comme signe de délimitation d'une unité lexicale simple. Les unités lexicales composés doivent être mis dans la base de données

LEXIQUE.

Nous reviendrons sur le traitement des signes de ponctuation, tiret et ponctuation double car elles font partie de la catégorie T du modèle proposé.

3 Etablissement de la Grammaire de référence

La procédure d'identification et d'extraction des syntagmes nominaux passe d'abord par la reconnaissance de leurs constituants, leurs unités lexicales. Or, la reconnaissance de ces constituants n'est pas une tâche facile, outre les ambiguïtés des unités lexicales dont nous avons parlé (la synonymie et la polysémie), il y a ambiguïté par rapport au rôle qu'elles peuvent jouer dans une phrase. Exemples de ces problèmes : les cas de quelques unités lexicales qui peuvent jouer le rôle d'un substantif (nom) dans un contexte et, d'adjectif dans un autre. De plus, dans la grammaire traditionnelle on trouve des mots classés dans une catégorie qui en fait, jouent le rôle d'une autre catégorie. Par exemple des participes utilisés souvent à la place d'un adjectif, des pronoms toniques qui occupent souvent la place d'un nom substantif.

M. LE GUERN soutiens que ***« La question des parties du discours se situe, le plus souvent de manière implicite, à l'articulation du lexique et de la syntaxe. Il s'agit, en effet, de déterminer les sous-ensembles du lexique contenant les éléments qui se voient assigner dans le discours le même comportement syntaxique. »*** [76]

Par ailleurs, comme en français, on trouve aussi dans la langue portugaise les articles contractés, c'est-à-dire les mots qui sont composé d'une préposition et d'un article (de + article, em + article ou por + article) :

da
de + a (de la) ;

do
de + o (du) ;

na
em + a (en la) ;

no
em + o (en le) ;

pelo
por + o (par le).

Le problème est de savoir comment classer ces unités lexicales ou plutôt comment les traiter. Où doit-on les classer si on utilise les catégories de la grammaire traditionnelle ? Est-ce qu'il faut créer une catégorie « ad-hoc », appelée Articles Contractés ? Or, ce genre de mots est formé à partir de deux autres mots qui appartient, en fait, à des catégories déjà existantes.

[76] *Michel LE GUERN. « Parties du discours et catégories morphologiques en analyse automatique ».* Les Classes de Mots. Lyon : Presses Universitaires de Lyon, 1994. p. 208.

La création d'une classe « ad-hoc » ***« …n'apporte et ni retranche aucune information structurale par rapport aux formes de surface … et ne peut donc constituer, pour l'analyse syntaxique, une meilleure base de départ que ces formes elles-mêmes. »*** [77] . A quoi sert cette catégorisation des unités lexicales ? La catégorisation des mots devrait prendre en compte le but du traitement de ces unités. Ce sont quelques arguments utilisés par Alain BERRENDONNER pour montrer le besoin d'avoir un système de classification, plus homogène et qui puisse servir non seulement à une tâche purement classificatoire, mais aussi à d'autres finalités comme celle de l'analyse syntaxique, de l'indexation automatique etc. Ainsi, Alain BERRENDONNER a établi deux principes de base pour la construction d'un analyseur morphologique : 1) ***« il faut commencer par définir explicitement et rigoureusement un ensemble de conditions auxquelles doit satisfaire son produit de sortie R, compte tenu des fonctions qu'on envisage de lui confier dans la suite du traitement. »*** [78] ***; 2) Le second principe est que définir ainsi un produit de sortie pour l'AM, c'est, du même coup, choisir une grammaire de référence : R la représentation visée est un certain type d'analyse linguistique du texte T, e formuler certaines exigences explicites à propos de R revient à sélectionner, parmi toutes les grammaires de références possibles, celle qui est capable de générer une représentation conforme à ces exigences. »*** [79] .

Ces deux principes peuvent être représentés par le schéma de la figure 9.3 (schéma emprunté à l'ouvrage d'Alain BERRENDONNER cité plus haut), montrant le rapport entre les éléments d'entrée et de sortie avec l'analyseur morphologique. Le résultat R est fonction du Texte T et de la grammaire de référence G.

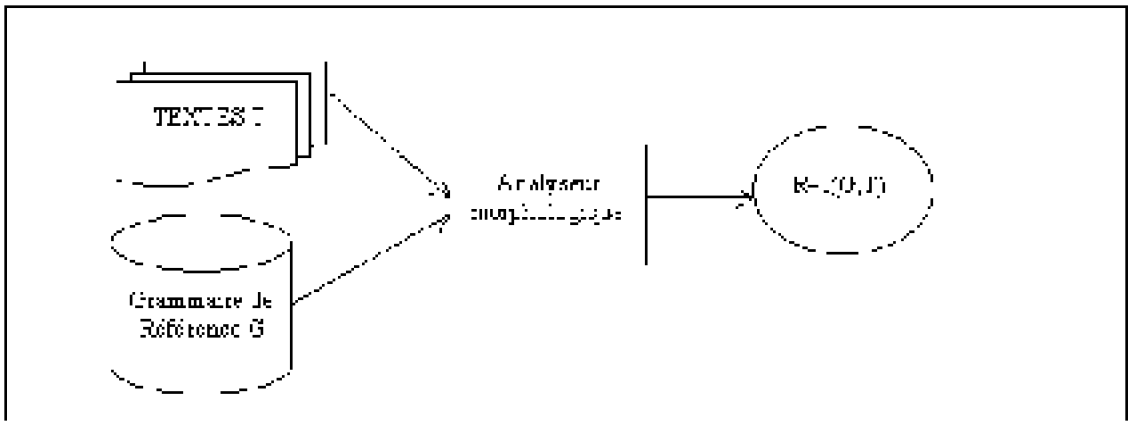
Ainsi, dans ce travail la grammaire de référence donnera la définition d'un ensemble de caractéristiques nécessaires à repérer, pour chaque unité lexicale (mot, mot composé), en vue de faciliter l'identification d'un syntagme nominal. Ces caractéristiques sont des informations spécifiques à chaque unité lexicale, comme sa catégorie grammaticale, sa flexion en genre, en nombre, sa personne parmi d'autres spécificités.

Comme il a été déjà signalé, nous allons utiliser comme base pour ce travail l'approche développée par Alain BERRENDONNER pour la langue française. D'abord, ses arguments nous semblent très pertinents et il a réussi à construire un modèle de classification beaucoup plus homogène, cohérent et enrichi que celui de la grammaire traditionnelle. Le produit de son travail est, donc beaucoup plus approprié au traitement automatique de textes, soit pour la reconnaissance et extraction des syntagmes nominaux, soit pour une analyse syntaxique.

[77] Alain BERRENDONNER. *Grammaire pour une analyseur :aspects morphologiques*. Les Cahiers du Criss. Grenoble :Centre de Recherche en Informatique appliquée aux Sciences Sociales. Université des Sciences Sociales de Grenoble, Novembre, 1990. p. 4.

[78] ***Ibidem p. 3.***

[79] Ibidem p. 3.



La langue portugaise ressemble beaucoup à la langue française, elles ont la même origine, leurs structures grammaticales sont pareilles. Cependant, comme toutes les langues, le portugais a quelques spécificités qui marquent une différence.

Le modèle proposé demande la régularisation de toutes formes d'amalgames à travers leur décomposition suivant leurs composants primitifs étant donné que ces composants sont déjà classés dans des catégories existantes dans la grammaire de référence. Ainsi, il faut faire un pré-traitement de ces unités lexicales, comme par exemple les articles contractés (ao, do, da, à, etc.), comme les amalgames préposition + pronom démonstratif (daquele de + aquele, naquele em + aquele, etc) parmi d'autres cas. Le pré-traitement peut être réalisé soit comme un premier balayage sur tout le texte soit au fur et à mesure qu'on traite chaque unité lexicale. La dernière option est plus intéressante du point de vue de vitesse de traitement, tandis que la première oblige à faire deux parcours sur le même texte, ce qui demande un peu plus de temps.

S'on choisit de faire le pré-traitement au moment de l'identification de chaque unité lexicale, il faut encore savoir comment le faire, on peut mettre les formes contractées dans la base de données LEXIQUE ou les déclarer dans le module de Segmentation et Identification d'Unités Lexicales (SIDUL). La première solution est, à notre avis, la meilleure, étant donné l'avantage de construire un module SIDUL plus indépendant des données lexicales. Le seul inconvénient est le temps de traitement qui peut rendre l'analyse un peu plus lente. Il est vrai que la deuxième voie serait une solution plus rapide en ce qui concerne le temps de traitement de données. Cependant, elle rend le module SIDUL plus particulier puisque toutes les contractions ou amalgames y sont mis. Il faudra, donc, le changer chaque fois qu'on aura besoin d'inclure une nouvelle forme de contraction. Ce qui peut rendre le logiciel d'analyse un peu gourmand par rapport à l'utilisation de la mémoire.

3.1 Les catégories et variables

La démarche choisie consiste à organiser et établir les catégories des unités lexicales selon leurs propriétés distributionnelles dans le discours. En fait, au début il y avait trois approches possibles, soit on utilisait les catégories existantes dans la grammaire traditionnelles (substantifs, adjectifs, pronoms, verbes, articles, etc.), soit on adaptait un modèle différent de celui de la grammaire traditionnelle (comme celle conçue par Alain

BERRENDONNER) soit on développait une autre grammaire. Or, le premier choix avait les inconvénients déjà cités dans la section précédente. Le troisième exigeait des connaissances linguistiques beaucoup plus approfondie et demanderait plus de temps. Ce n'était pas possible de le réaliser dans le cadre de cette recherche, cette troisième voie pourrait être objet d'une autre thèse, telle est la complexité. Il était hors question de la prendre. La meilleure solution était donc de faire une adaptation de l'approche proposée par Alain BERRENDONNER pour le portugais.

Pour une question de compatibilité avec le modèle de classification conçu pour la langue française, on cherchera à utiliser la même notation, aussi bien pour les catégories que pour les variables et leurs valeurs. Nous avons adopté quand même quelques valeurs particulières pour celles déjà prises pour la langue française étant donné soit la spécificité de la langue portugaise, soit le but de ce modèle : reconnaître et extraire les SN.

Ainsi, la catégorie majeure sera désignée par un caractère, les variables de sous-catégorisation seront désignées par deux caractères et finalement les valeurs possibles pour chacune des variables de sous-catégorisation seront indiquées par trois caractères. La valeur non marquée sera toujours le nom de la variable de sous-catégorisation plus une lettre 'N' à la fin.

Pour le pré-traitement des amalgames, nous avons adopté d'indiquer cette procédure à travers une variable appelée RG (règles), et la valeur appelée PRE. Cependant, l'indication du pré-traitement, n'est pas suffisant, il faut indiquer aussi respectivement quels sont leurs éléments primaires et leur catégorie. Ces indications seront faites dans une relation ou table grâce à la structures de la base de données LEXIQUE. Nous allons adopter cette procédure parce que le module d'analyse SIDUL doit analyser les constituants primaires, connaissant d'abord leurs caractéristiques. En fait, PRE est une des valeurs possibles pour la variable appelé RG. Cette variable doit comprendre d'autres règles, qui seront discutées au fur et à mesure qu'elles sont crées dans ce chapitre.

Exemple : *ao* => *a* + *o*=> (ao, PRE, a, P, o, D) [au => à + le => (ao, PRE, à, P, le, D)]

Dans cet exemple, on interprète les paramètres présents entre les parenthèses comme étant une unité lexicale qui est constituée par une contraction d'une préposition 'a' avec un article 'o', où le code PRE indique qu'il faut faire un pré-traitement, c'est-à-dire la décomposition de 'ao' en 'a' P, préposition, et 'o' D, déterminant.

On montrera à la fin de ce chapitre la structure de la base de données LEXIQUE.

Ainsi que dans la langue française, les unités lexicales du vocabulaire de la langue portugaise seront reparties en 10 (dix) classes majeures. Le modèle adopté par Alain BERRENDONNER est basé ***« sur les propriétés distributionnelles considérées comme principales (=celles qui caractérisent le comportement d'un mot au regard de la syntaxe de constituants). »*** [80] .

[80] Alain BERRENDONNER. Grammaire pour une analyseur :aspects morphologiques. Les Cahiers du Criss. Grenoble :Centre de Recherche en Informatique appliquée aux Sciences Sociales. Université des Sciences Sociales de Grenoble, Novembre, 1990. p. 18.

Cependant les catégories ne sont pas suffisantes pour représenter, elles seules, la totalité du comportement combinatoire d'un mot. Ainsi ; ce modèle est composé, outre les catégories majeures, de quatre types de variables, appelées variables de sous-catégorisation :

1. variables de sous-catégorisation syntaxique — ce sont des variables pour indiquer plus précisément le comportement d'une unité lexicale en tant que constituant syntaxique. Exemple de ce type de variable est celle nécessaire pour indiquer la structure ternaire caractéristique des nominaux en français et aussi en portugais. Des mots peuvent jouer le rôle de nom, d'autres jouent le rôle d'adjectif tandis que d'autres peuvent jouer le rôle autant de nom que d'adjectif. Pour cela, bien qu'on les classe tous dans la catégorie F, il faut une sous-catégorisation pour refléter plus précisément cette particularité. Ainsi, NA est une variable appelée type nominaux, qui indique pour chaque mot le rôle qu'il peut jouer dans la phrase (NOM, ADJ, NAN), c'est-à-dire nom, adjectif ou non marqué ;
2. variables de sous-catégorisation flexionnelles — comme le nom de ce genre de variables l'indique, elles contiennent des informations sur la flexion d'une unité lexicale (genre, nombre et personne) ;
3. variables de sous-catégorisation lexicale — ces variables contiendront des informations sur « les principales restrictions portantes sur la cooccurrence d'un item lexicale avec d'autres items (restrictions sélectives, rections valences etc.). »Alain BERRENDONNER. Grammaire pour une analyseur :aspects morphologiques. Les Cahiers du Criss. Grenoble :Centre de Recherche en Informatique appliquée aux Sciences Sociales. Université des Sciences Sociales de Grenoble, Novembre, 1990. p. 19. ;
4. variables de caractère générique — nous allons créer une sorte de variables pour garder des règles qui seront établies de manière à résoudre des problèmes impossibles en utilisant simplement les autres variables. Ces règles seront définies au fur et à mesure qu'on trouve et discute chaque problème.

Nous allons décrire maintenant chaque catégorie et ses particularités en les discutant et établissant les solutions possibles pour chaque cas. Cette discussion sera menée en ayant le corpus de syntagmes nominaux extraits dans le cadre de notre mémoire de DEA comme base. Dans certains cas, nous avons pris des exemples dans la grammaire de Celso CUNHA & Lindley CINTRA. Il faut dire que des constructions syntaxiques plus courantes dans la littérature et dans la poésie ne sont guère utilisées dans le discours technique. Dans ce sens là, ce travail n'est pas exhaustif.

3.1.1 Verbes

Catégorie représentée par le caractère « V ». ***« Les verbes sont des mots qui peuvent présenter, morphologiquement, de variations en nombre, en personne, en mode, en temps, en aspect et en voix. »*** [81] . En Portugais, du point de vue distributionnelle, le contexte caractéristique, des verbes, serait tous ceux qui peuvent suivre immédiatement la particule de négation, dans la forme :

X Não __ Y

La différence entre le contexte distributionnelle dans le portugais et dans le français, est qu'en français la négation est faite de manière presque souvent avec deux particules, tandis qu'en portugais il n'y en a qu'une seule particule.

De la même façon que le contexte distributionnel pour le français, ce contexte recouvre ce qui est traité comme forme verbale, y compris les infinitifs. Cependant il exclue le participe. Il faut remarquer qu'en Portugais il n'existe qu'une forme du participe. Il n'y a pas de participe passé ou présent. En fait, le gérondif (« gerundio ») ressemble au participe présent qui peut apparaître dans une phrase avec le rôle d'adjectif. Etant donné que les formes verbales au participe jouent le rôle d'adjectif et peuvent être fléchies en genre et nombre ils seront mis dans la catégorie « F ».

Pourtant, lorsque le participe est précédé par un verbe auxiliaire comme *ter* (avoir)*, haver* (avoir)*, estar*(être) ou *ser* (être) il joue le rôle d'un verbe. Ce sont de formes composés du verbe. Exemples :

1. ter escrito (avoir écrit) ;
2. Estamos impressionados com a situação (nous sommes impressionés par la situation) ;
3. A carta foi escrita por mim. (La lettre a été écrite pour moi.).

Ainsi, dans le cadre de cette catégorie, pour faciliter la reconnaissance des formes verbales, on va créer aussi une variable pour distinguer les formes verbales, celle qui on appellera VX.

VX =
{AUX, ORD} oùAUX indique que l'unité lexicale est un verbe auxiliaire ;ORD indique que l'unité est un verbe non auxiliaire.

Il faut aussi avoir une variable pour indiquer les mots dont la forme du participe ou du gérondif et pour identifier les situations où ils composent des formes verbales composes. Etant donné qu'elles jouent souvent le rôle d'adjectifs, elles sont mises dans la catégorie F. On a donc la variable PG pour désigner les unités lexicales qui sont dans la forme du participe ou du gérondif :

PG =
{PAR, GER} où :PAR désigne ce qui est dans la forme verbale au participe ;GER désigne ce qui est dans la forme verbale au gérondif.

RG =
{VSV} oùVSV indique à l'analyseur que l'unité sera interprétée comme une forme verbale s'elle vient après un verbe auxiliaire.

Une autre variable importante pour identifier les mots dans la forme verbale est celle

[81] Celso CUNHA & Lindsey CINTRA. *Nova gramática do português contemporâneo.* Lisboa : Edições João Sa da Costa, 1984. p. 377.

indiquant si la forme est à l'infinitif ou finie.

VB =
{INF, FIN} où :INF indique que la forme verbale est à l'infinitif ;FIN indique que la forme verbale est à la forme finie.

La forme à l'infinitif du verbe peut jouer le rôle de nom substantif s'il est précédé par un déterminant. Exemple : O nascer do dia (la naissance du jour).

Ainsi, on va créer une sorte de variable pour signaler cette caractéristique. En fait cette variable contiendra des valeurs qui ont le rôle de règles de contraintes.

RG =
{ VSV, NSP } où

NSP indique au système analyseur que s'il y a un déterminant (D) précédant un verbe, ce dernier est alors dans le rôle d'un nom substantif, c'est à dire, il peut être considéré comme un élément de la catégorie F. Cette variable, RG, possède d'autres valeurs lesquelles nous expliciterons au fur et à mesure des besoins tout au long de ce travail.

3.1.1.1 Rection verbale

On a vu dans la description de l'expérimentation de la maquette du système de recherche d'information, construit dans le cadre du mémoire de DEA, que la double rection était la raison d'une des difficultés de structuration des syntagmes nominaux. À ce moment là, le seul problème rencontré a été la difficulté d'établir le rapport entre un SN et un autre suivant leur structuration, parce que l'extraction a été déjà faite, de manière « artisanale », par nous- même. C'est-à-dire de façon non automatisée. Mais pour l'identification et pour l'extraction automatique des SN il faut faire attention à la double rection, car si on ne tient pas compte de la double rection, on extrait des faux syntagmes nominaux.

Exemple :

1. a biblioteca fornece informações cientificas aos usuários (la bibliothèque fourni des informations scientifiques aux usagers) L'extraction des SN sans observer la double rection donnera les syntagmes nominaux suivants ;

1. a biblioteca (la bibliothèque) [SN de niveau 1]
2. informações cientificas aos usuarios (des informations scientifiques aux usagers) [SN de niveau 2]
3. os usuarios (les usagers) [SN de niveau 1] Cependant, il y a une double rection, le verbe forcencer (fournir) est un verbe qui peut avoir jusqu'à deux compléments, un complément d'objet direct et l'autre d'objet indirect. Ainsi, les syntagmes corrects devrait être :
4. a biblioteca (la bibliothèque) [SN 1]
5. informações científicas (des informations scientifiques) [SN de niveau 1]
6. os usuarios (les usagers) [SN de niveau 1] Par extraction des SN en observant la

double rection, on obtient deux SN indépendants (e, f) en opposition à ceux de la première extraction (b, c) lesquels sont enchaînés étant donné que le SN (a) est de deuxième niveau. D'ailleurs, du point de vue logico-sémantique, cette forme n'est guère acceptable.

La conséquence de l'extraction des faux SN est reflété dans la structure des SN et aussi dans la recherche d'information. Si on prend l'exemple donné :

a biblioteca forncece informações cientificas aos usuarios

On n'arrive pas au SN *informações científicas aos usuários* à travers le centre de SN de premier niveau *informações*, car ce mot n'est pas le centre du SN de premier niveau, il est centre mais d'un SN de deuxième niveau. Ainsi, en utilisant la maquette développée dans le cadre du DEA, il fallait appeler le SN de premier niveau, par le biais du mot *usuários,* qui est le centre du SN de premier niveau *os usuarios.* Tandis que, si on arrive à extraire correctement ces SN, on a les SN *informações científicas* et *os usuários.* Ce qui permet d'accéder au SN *informações científicas,* par le centre *informações* et également au SN *os usuários,* à travers le centre de SN *usuários.*

Ainsi, il faut prendre en compte les compléments exigés par les verbes et aussi, bien distinguer les syntagmes nominaux qu'y sont présents comme des SN distincts. C'est-à-dire, extraire de manière indépendante chaque SN présent dans chaque complément.

Nous allons définir une variable pour aider la procédure d'identification et d'extraction des SN pour résoudre ce problème. Le but de cette variable, qu'on appellera NC, nombre de compléments, est d'indiquer le nombre de compléments demandé par la forme verbale. Cette variable ressemble un peu à celle définie par Alain BERRENDONNER (VA - valence). La variable VA indique le nombre d'actants tandis que la variable NC indique le nombre de compléments demandés par la forme verbale. C'est-à-dire qu'on ne compte pas la quantité d'actants d'un verbe, mais la quantité de compléments maximaux qu'un verbe peut exiger.

NC =
{0CO,1CO, 2CO, 3CO,4CO} où :0CO pour dire que le verbe n'exige aucun complément ;1CO pour dire que le verbe exige jusqu'à un complément ;2CO idem pour jusqu'à deux compléments ;3CO idem pour jusqu'à trois compléments ;4CO idem pour jusqu'à quatre compléments, ce qui est très rare.

De plus, il faut informer le système analyseur, des prépositions qui peuvent précéder chaque complément. C'est-à-dire, il faut informer les combinaisons complètes de prépositions qui précèdent les compléments pour chaque verbe. Pour cela, dans la base de donné LEXIQUE il est prévu une table (PREP) où seront enregistré chaque combinaison de préposition ayant le verbe comme clé primaire. Les formes verbales transitif direct n'exige pas des prépositions, en ce cas il n'y aura aucune indication de prépositions dans la table correspondant dans la base de données LEXIQUE, mais un simple espace.

Bien que cette solution nous semble suffisante puisqu'elle peut résoudre un nombre

important de cas de rection, elle peut échouer cependant devant deux facteurs de risque : a) une mauvaise utilisation des prépositions par l'auteur ; b) l'apparition de compléments inattendus comme des compléments de nominalisations ou des compléments circonstanciels.

En ce que concerne la mauvaise utilisation des prépositions par l'auteur du document, actuellement on ne peut rien faire, la conséquence peut être le manque de SN ou plus probablement l'indexation des faux SN dans leur structure. Ce qui peut augmenter le nombre de faux SN et, en conséquence le taux de silence par rapport à certains SN dans la procédure de recherche d'information. Une solution pour résoudre ce problème serait corriger la préposition, à travers la correction grammaticale automatique avant de faire le traitement du texte. Mais cela demande la construction d'un analyseur syntaxique automatique.

Le deuxième point (b), il est plus complexe car il faut trouver des moyens de reconnaître tous les compléments. Les compléments dus à la rection des mots, soit des verbes, soit des noms ou des adjectifs, peuvent être identifiés en utilisant la variable NC de façon successive. Or, même en faisant ce genre de traitement, on risque encore d'avoir des problèmes car ce sont des cas où on trouve un mot qui n'exige pas de compléments. C'est-à-dire, un mot caractérisé avec NC=0CO, en précédant un syntagme prépositionnel ou même une expansion prépositionnelle. Ici l'ennui arrive lorsque la préposition qui appartient à ce dernier syntagme prépositionnel ou à cette dernière expansion prépositionnelle est la même qui régit le deuxième complément du verbe. Empruntons l'exemple donné par Jean-Paul METZGER.

la vente à la cantine de produits frais Selon la solution adopté on a : NC = 2CO Prépositions : à, de SN_i : *la cantine* SN_{ii} : *des produits frais*

Par contre, si l'on prend l'exemple donné par Omar LAROUK, en changeant juste l'ordre de la rection :

la vente à la cantine de l'université de produits frais

Selon la solution adoptée sans se rendre compte de l'existence du complément de la cantine. Cela donne :

NC = 2CO Prépositions : à, de SN_i : *la cantine* SN_{ii} : *l'université de produits frais*

Le deuxième SN est faux et l'extraction correcte devrait être : SN_i : *la cantine de l'université* SN_{ii} : *des produits frais*

Il est vrai que cet exemple est fabriqué et que ce SN serait rédigé vraisemblablement, dans le quotidien, de façon directe (*la vente de produits frais à la cantine de l'université)*. Pourtant, il sert à montrer que la procédure choisie pour reconnaître et extraire les SN dans des cas de rections multiples peut échouer dans certains cas. Il s'agit bien sûr d'une construction peut courante mais qui peu se rencontrer. De toute façon, une analyse multiple serait envisageable, en accord à M. LE GUERN, malgré la multitude de SN résultats d'une telle analyse, on est sûr que le bon SN serait extrait. Ce genre d'analyse va sans doute fournir de faux SN, ce qui n'est pas trop grave car le but de ce modèle est l'indexation automatique et non pas la traduction automatique ou l'analyse syntaxique complète. Il est vrai qu'on peut avoir une croissance de bruit dans la structure de SN. Or,

le système de recherche proposé a la particularité d'être un système assisté par ordinateur, où le point fort est l'interactivité avec l'usager. C'est à lui de bien choisir. Dans ce sens là, il faut traiter les cas les plus courants au lieu de créer de multiples variables pour résoudre des cas moins courants, ce qui va certainement alourdir le système.

L'exemple que nous venons de fabriquer est une sorte de nominalisation. C'est un problème difficile à résoudre, ainsi que les problèmes des compléments circonstanciels. Du point de vue théorique, on peut les comprends à l'aide d'un système causal. Mais il est insuffisant pour un traitement automatique, car il n'y a guère des marques qui puissent permettre un ordinateur de les reconnaître et de les traiter de forme automatique. Malheureusement l'ordinateur est aveugle et destitué de capacité de raisonnement, ce sont les humains qui lui donnent une certaine capacité de prise de décision, une certaine capacité de raisonnement, par le biais des logiciels et aussi des fichiers remplis de connaissances spécifiques.

Dans un autre côté, il existe des études en cours sur la nominalisation, il faut donc attendre les résultats de celles-là pour en tirer parti et améliorer l'analyseur que nous développons actuellement.

3.1.2 Nominaux

Catégorie représentée par le symbole « F ». Cette catégorie de mots regroupe les noms. Antoine ARNAULD & Claude LANCELOT montrent l'origine de cette classe de mots.

« Les objets de nos pensées sont ou les choses, comme la terre, le soleil, l'eau, le bois, ce qu'on appelle ordinairement substance ; ou la manière des choses, comme d'être rond, d'être rouge, d'être dur, d'être savant, etc., ce qu'on appelle accident. « Et il y a cette différence entre les choses et les substances, et la manière des choses ou des accidents, que les substances subsistent par elles-mêmes, au lieu que les accidents ne sont que par les substances. « C'est ce qui a fait la principale différence entre les mots qui signifient les objets des pensées : car ceux qui signifient les substances ont été appelés noms substantifs ; et ceux qui signifient les accidents, en marquant le sujet auquel ces accidents conviennent, noms adjectifs. » [82]

Ce qu'on peut retenir de cette citation, c'est qu'à l'origine il y avait déjà une classe nominale avec une sous-catégorisation pour distinguer les noms substantifs et les noms adjectifs. Un peu plus loin dans la Grammaire Générale et Raisonnée de Port-Royal, les auteurs montrent aussi que les substantifs sont des noms qui subsistent par eux-mêmes dans les discours, sans avoir besoin d'un autre nom, bien qu'ils signifient des accidents. Au contraire, en opposition, les noms adjectifs, sont des noms qui doivent être joints à d'autres noms dans le discours.

Une autre constatation faite par Antoine ARNAULD & Claude LANCELOT est que même des noms qui désignent des accidents, c'est-à-dire appartiennent à la catégorie des noms adjectifs, peuvent subsister tout seul comme des noms substantifs, ce sont des noms de diverses professions, comme : artiste, philosophe, peintre, soldat, etc. Selon

[82] ***Antoine ARNAULD & Claude LANCELOT. Grammaire Générale et Raiosonnée de Port-Royal. Généve : Slatkine Reprints, 1993. p. 48-49.***

eux : ***« Et ce qui fait que ces noms passent pour substantifs, est que pouvant avoir pour sujet que l'homme seul, au moins pour l'ordinaire, et selon la première imposition des noms, il n'a pas été nécessaire d'y joindre leur substantif, parce qu'on l'y peut sous-entendre sans aucune confusion, le rapport ne s'en pouvant faire à aucune autre ; et par-là ces mots ont eu dans l'usage ce qui est particulier aux substantifs, qui est de subsister seuls dans le discours.»*** [83] .

Ainsi, cette citation vient corroborer la démarche suivie par Alain BERRENDONNER, lorsqu'il met les substantifs et les adjectifs dans une même catégorie, appelée « F ». Celso CUNHA & Lindley CINTRA montre aussi qu'on trouve le même phénomène dans la langue portugaise :

« Le rapport entre le substantif (terme déterminé) et l'adjectif (terme déterminant) est trop étroit. Ce n'est pas rare, il y a une forme unique pour les deux classes de mots et, dans ce cas, la distinction ne peut être faite que dans la phrase. Comparez, par exemple : Uma preta velha vendia laranjas. (Une noire vieille vendait des oranges) Uma velha preta vendia laranjas. (Une vieille noire vendait des oranges) « Dans la première phrase, preta (noire) est substantif, car elle est le mot noyau caractérisé par velha (vieille), qu'à son tour est adjectif dans la mesure qu'il fait la caractérisation du terme noyau. Dans la deuxième phrase, au contraire, velha (vieille) est le substantif et preta (noire) l'adjectif. » [84]

Ils arrivent à la conclusion que la distinction entre substantifs et adjectifs obéit à un critère purement syntaxique, fonctionnel. Il me semble que dans ce cas, on peut résoudre le problème en tenant compte de la proximité du déterminant, c'est-à-dire, en considérant comme substantif le terme qui est le plus proche du déterminant.

Nous avons montré que le problème de classification de mots dans la catégorie des nominaux en portugais ressemble à celui du français. Le contexte établi pour caractériser les noms e adjectifs dans la langue française peut être établi aussi pour le portugais :

« Ce(s) être D_X » ou « Ce(s) être D_ » (« Este(s) ser D_X » ou « Este(s) ser D_ ») 1.

« Il(s) être _X » ou « Il(s) être _ » (« Ele(s) ser_X » ou « Ele(s) ser _ ») 2.

Où « D » représente un élément de la catégorie des prédéterminants. « X » est une éventuelle séquence régie par l'élément nominal à tester. Exemples :

Esta é uma velha(a) - (C'est une vieille) 1.

[83] Antoine ARNAULD & Claude LANCELOT. *Grammaire Générale et Raiosonnée de Port-Royal.* Genéve : Slatkine Reprints, 1993. p. 50.

[84] « É muito estreita a relação entre o substantivo (termo determinado) e o adjectivo (termo determinante). Não raro, ha uma única forma para as duas classes de palavras e, nesse caso, a distinção só podera ser feita na frase. Comparem-se por exemplo : Uma preta velha vendia laranjas. Uma velha preta vendia laranjas. Na primeira oração, preta é substantivo, porque é a palavra-núcleo, caracterizada por velha, que por sua vez, é adjectivo na medida em que é a palavra caracterizadora do termo-núcleo. Na segunda oração, ao contrario, velha é substantivo e preta adjectivo. » citation extrait de Celso CUNHA & Lindsey CINTRA. Nova gramática do português contemporâneo. Lisboa : Edições João Sa da Costa, 1984. p. 248.

Ela é velha(b) - (Elle est vieille) 2.

Esta é uma negra (a) - (C'est une femme noire) 3.

Ela é negra(b) - (Elle est femme noire) 4.

Les exemples corroborent les exemples donnés par Celso CUNHA & Lindley CINTRA, parce que les quatre formes sont acceptables, *velha* et *negra* peuvent jouer le rôle de substantif ou d'adjectif dépendant du contexte. Or, l'exemple d'application fonctionne avec les exemples en portugais, nous ne sommes pas sûr que les mêmes exemples puissent fonctionner en français car la façon d'exprimer les choses en portugais ne sont pas toujours comme en français.

Cependant, d'autres mots ne jouent qu'un seul rôle, soit substantif, soit adjectif.

Exemples :

Este é um sistema(a) - (C'est un système) 1.

Ele é sistema (b) - (Il est système) 2.

Este é um substancial (a) - (C'est un essentiel) 3.

Ele é substancial(b) - (Il est essentiel) 4.

Les exemples (5) et (8) sont des phrases acceptables, tandis que les phrases (6) et (7) ne le sont pas. Il s'agit de mots qui jouent un seul rôle, c'est-à-dire, sistema (système) est un substantif alors que substancial (essentiel) est un adjectif.

Ainsi, il faut créer une variable de sous-catégorisation NA pour indiquer plus précisément le rôle qu'un mot peut jouer dans un syntagme nominal. Ainsi, cette variable peut avoir les valeurs suivantes :

NA =
{NOM, ADJ, NAN} où :NOM pour désigner les mots qui jouent seulement le rôle d'un nom, d'un substantif ;ADJ pour désigner les mots qui jouent seulement le rôle d'un adjectif ;NAN pour désigner les mots qui sont non marqués, ils peuvent être soit un nom, soit un adjectif.

Il faut aussi distinguer les formes fléchies des mots, en genre et en nombre, ce qui implique deux autres variables :

NB =
{PLU, SIN, NBN} où :PLU pour désigner les mots qui sont au pluriel ;SIN pour désigner les mots qui sont au singulier ;NBN pour désigner les mots qui sont non marqués en nombre.

GR =
{MAS, FEM, GRN} où :MAS pour désigner les mots masculins ;FEM pour désigner les mots féminins ;GRN pour désigner les mots non marqués en genre.

D'autres variables sont encore nécessaires pour l'identification des SN. Nous avons vu dans l'étude sur l'omission d'articles devant quelques syntagmes nominaux, dans le chapitre précédant qu'une des marques souhaitables pour être repérée est de savoir s'il

est concret ou abstrait.

VN =
{CON, ABS} oùCON pour indiquer qu'un mot est un nom concret ;ABS pour indiquer qu'un mot est un nom abstrait.

Nous sommes tentés, tout d'abord, d'adopter les traits matériels et immatériels au lieu d'adopter les traits concrets et abstraits. Cependant, il nous semble que la valeur abstraite est plus appropriée car le trait d'immatériel n'est qu'une des acceptions du terme « abstrait ». Selon Marc WILMET, ***« Le constat vaut aussi des NA, avec la circonstance aggravante que la polysémie du verbe abstraire et de l'adjectif abstrait a produit au moins sept acceptions concurrentes. »*** [85] . Selon lui ces acceptions sont : extrait, immatériel, général, conceptuel, abscons, réduit ou subduit, dérivé. Ainsi nous avons résolu de laisser comme valeur pour cette variable CON et ABS.

En ce qui concerne les adjectifs, il faut repérer une caractéristique spécifique de ces unités. Nous allons montrer dans le chapitre prochain, le rôle des adjectifs de relation dans la syntaxe du syntagme nominal. Ces genres d'adjectifs cache, à l'intérieur, un syntagme prépositionnel. Ce qui leur donne un caractère d'adjectif complexe. Normalement, un N + A donne un N. Pourtant, un N + A' ne donne pas un N, mais un N' puisque l'adjectif complexe contient un syntagme prépositionnel à l'intérieur. C'est-à-dire :

Exemple : Le café brésilien (D' + N + A') où A' est un adjectif de relation.

« brésilien » est un adjectif de relation qui peut être remplacer par « du Brésil ». Le syntagme présenté dans l'exemple peut donc être réécrit comme « Le café du Brésil ». L'adjectif de relation « brésilien » contient donc un syntagme prépositionnel (« du Brésil »). La réécriture de ce syntagme est :

D' + N + P' + D' + N

Si nous réduisons cette description, étant donné que D' + N peut être réécrit comme D' + N' et finalement comme N'', la description peut être réécrit comme :

D' + N + P' + N''

Or, P' + N'' est un syntagme prépositionnel, donc un SP :

D' + N + SP

Et N + SP peut être réécrit comme N' :

$(D' + N')_{N''}$

Ainsi, nous avons établi une autre variable pour caractériser les mots qui peuvent jouer le rôle d'un adjectif de relation :

TA =
{QUA, REL} oùQUA pour les adjectifs de qualité ;REL pour distinguer les adjectifs de relation.

[85] Marc WILMET. « A la recherche du nom abstrait ». In. : Nelly FLAUX, Michel GLATIGNY et Didier SAMAIN. *Les Noms Abstraits : histoire et théorie*. Collection Sens et Structures. Paris : Presses Universitaires du Septentrion, 1996. p. 67.

Nous avons dit que « les mots qui peuvent jouer le rôle d'un adjectif de relation » parce que ces adjectifs peut aussi jouer le rôle d'un adjectif de qualité selon le contexte. Nous reviendrons à cette discussion dans la grammaire de reconnaissance et d'extraction des SN.

Bien que nous n'allions pas traiter ici des résolutions des éléments anaphoriques, il est prudent de prévoir quelques variables qui pourront être utilisés par des futurs travaux relatifs à la résolution de ces éléments. C'est pourquoi nous envisageons déjà deux variables : a) une variable utilisée pour sous-classer les unités F et Y en fonction de l'axe animé / inanimé ; b) une autre variable pour indiquer la flexion en personne de quelques-unes unes des unités classées comme F et Y. Ainsi :

AN =
{ ANI, INA, ANN } oùANI désigne ce qui est animé ;INA désigne ce qui est inanimé ;

ANN pour indiquer le caractère non marqué par rapport au trait animation.

En ce qui concerne la flexion en personne, nous avons adopté la même démarche utilisée par Alain BERRENDONNER [86] que s'est basé sur un travail de Émile BENVENISTE prouvant que l'utilisation de la flexion en nombre pour opposer le pronom personnel « *eu* » (je) et « *nós* » (nous), *tu* (tu) et vós (vous) est insuffisant. Émile BENVENISTE en conclut en disant : ***« La distinction ordinaire de singulier et de pluriel doit être sinon remplacée, au moins interprétée, dans l'ordre de la personne, par une distinction entre personne strict ('singulier') et personne amplifiée (= 'pluriel') Seule la troisième personne, étant non-personne, admet un véritable pluriel. »*** [87] . Les pronoms personnels (*eu* [je], *tu* [tu], *ele* [il], *ela* [elle], *nós* [nous], *vós* [vous], *eles* [ils], *elas* [elles]), sont employés de la même façon que leur respectifs en français. La seule différence est qu'en portugais il n'y a pas d'équivalent au pronom « on ». Mais les observations de Émile BENVENISTE sont aussi valables pour le portugais. C'est pourquoi nous allons adopter la même variable et leurs valeurs adoptées pour le français :

PE =
{PE1, PE2, PE3, PE4, PE5 } où :PE1 indique la première personne du singulier, cas de eu (je) ;PE2 indique la deuxième personne du singulier, tu (tu) ;PE3 indique la troisième personne, cas de ele, ela, eles, elas (il, elle, ils, elles) ;PE4 indique la première personne du pluriel, cas de nós (nous) ;PE5 indique la deuxième personne du pluriel, cas de vós (vous).

3.1.2.1 Autres formes nominales

[86] Alain BERRENDONNER. *Grammaire pour une analyseur :aspects morphologiques.* Les Cahiers du Criss. Grenoble :Centre de Recherche en Informatique appliquée aux Sciences Sociales. Université des Sciences Sociales de Grenoble, Novembre, 1990. p. 34.

[87] Émile BENVENISTE. *Problèmes de linguistique générale, 1.* Collection TEL. Éditions Gallimard, 1966. p. 235-236.

Selon Celso CUNHA & Lindsey CINTRA, en portugais il y a trois formes nominales du verbe. L'infinitif, le gérondif (*gerúndio* en portugais, équivaut au participe présent en Français) et le participe (équivaut au participe passé en Français). Leurs caractéristiques sont de n'exprimer par soi même ni le temps ni le mode [88] . Le verbe à la forme du participe peut fait partie de la catégorie F, parce qu'il peut jouer le rôle d'adjectif lorsqu'il n'exprime qu'un état, sans établir aucun rapport temporal [89] . Exemples :

1. a comunidade envolvida em pesquisa cientifica (la communauté impliquée dans la recherche scientifique)
2. a importância atribuída as pmi's (l'importance attribuée aux pmi's)
3. a informação aplicada (l'information appliquée)
4. a aceitação generalizada do conceito de consultoria informatológica (l'acceptation généralisée du concept de consultation informatologique)

Ce sont quelques exemples de mots dont le participe jouent ici le rôle d'adjectif.

En ce qui concerne l'infinitif non fléchi, il peut jouer le rôle d'un complément nominal après une préposition. Exemples :

1. a necessidade de ampliar a educação e o treinamento para a qualidade (le besoin d'amplifier l'éducation et l'entraînement pour la qualité)
2. a pretensão de isolar o sistema do homem (la prétention d'isoler le système de l'homme)
3. a capacidade de manipular a informação (la capacité de manipuler l'information)

L'infinitif peut aussi jouer un rôle de nom substantif, pour cela il faut qu'il soit précédé par un article ou un autre prédéterminant.

Un autre ensemble de mots méritent d'entrer dans la catégorie F, ceux qu'on appelle en portugais les pronoms « *obliquos* » dans la forme tonique (*mim, ti, ele, ela, nos, vos, eles, elas*), parce qu'ils jouent le rôle d'un substantif, ils ont le statut d'un syntagme nominal. En fait, ils font référence à une ou plusieurs personnes (*eles, elas, nos, vos*). Ils sont des anaphores, car ils font référence à des personnes ou des objets déjà cités dans le discours. C'est l'équivalent des pronoms toniques en Français (moi, toi, lui, eux, elles, nous, vous). Ces pronoms n'admettent pas de prédéterminant. D'ailleurs, quelques noms propres n'admettent pas non plus de prédéterminant comme précédemment.

Lorsque les unités, *mim, ti, nos, vos* sont placées après une préposition « *com* » (avec), on fait respectivement la contraction « *comigo, contigo, conosco, convosco* ». C'est-à-dire, il faut faire un pré-traitement. Pour cela, on adopte une variable indiquant le besoin de faire un pré-traitement avant de faire l'analyse. Cette variable appelée RG dont le but est de stocker des valeurs qui ont la force d'une règle.

[88] Celso CUNHA & Lindsey CINTRA. *Nova gramática do português contemporâneo.* Lisboa : Edições João Sa da Costa, 1984. p. 481.

[89] Ibidem p. 493.

RG =

{ VSV, NSP, PRE } où :PRE indique au SIDUL, module de segmentation et d'identification des unités lexicales, qu'il faut faire un pré-traitement de l'unité en cours d'analyse. Pour cela, il faut convertir la forme en cour, en deux autres, qui sont explicitées dans une table des amalgames. Cette table, a comme clé primaire l'unité lexicale contractée, et ses attributs sont, en fait, les éléments constituants de l'unité suivie de leur catégorie. Exemple : contigo = (com, P ; ti, F).

Parmi les pronoms indéfinis qui doivent faire partie de la catégorie F, on a : *alguém* (quelqu'un)*, ninguém* (personne)*, algo (quelque chose), nada* (rien). Pour illustrer l'utilisation de ces mots comme noms, nous empruntons à Celso CUNHA & Lindsey CINTRA les exemples suivants :

1. Minha Teresa tem algo a me dizer, não é ?Jorge Amado. Teresa Baptista cansada de guerra. São Paulo : Martins, 1972. p. 281. (Ma Thérèse a quelque chose à me dire, n'est-ce pas ?)
2. Ninguém ainda inventou fósforos contra o vento ?Augusto Abelaira. Quatro paredes nuas. Amadora : Bertrand, 1972. p. 25. (Persone n'a encore inventé des allumettes contre le vent ?)
3. E se alguém fosse avisar a Guarda ?Miguel Torga. Novos Contos da Montanha. 3ème edition. Coimbra : s.n., 1952. p.52. (Et si quelqu'un allait appeler la police ?)
4. Não devo nada a ninguémAlves Redol. Barracos de cegos. 4ème Edition. Lisboa :Europa-América, 1973. p.43..(Je ne doit rien à personne)

En ce que concerne le pronom *algo*, on peut le considérer comme un amalgame du pronom *alguma* + *coisa*, où coisa est F, NOM, COM [90] , SIN, FEM. En ce cas, il faut utiliser encore la variable RG avec la valeur PRE pour indiquer ce pré-traitement.

Ces pronoms sont aussi appelés, en grammaire portugaise, pronoms substantifs. Ils ont le même rôle que leurs correspondants en Français. De plus, d'autres pronoms indéfinis doivent être classés dans cette catégorie, soit : todo, todos, algum, alguns, alguém, outro, um.

Exemples :

1. Todos estavam admiradosCastro Soromenho. Terra morta. Lisboa : Sá da Costa, sans date. p. 186.. (Tous étaient étonnés.)
2. … nenhum teve para o outro a minima palavraRaul Pompéia. O Atheneu :chronica de saudades. 4ème Edition. Rio de Janeiro : Francisco Alves, sans date. p.205.. (…personne n'a eu pour l'autre le moindre mot)

Le mot « um » peut jouer le rôle d'un nom lorsqu'il est utilisé avec le prédéterminant

[90] COM appartient à un des valeur de la variable NN qui indique le type de nominaux, en ce cas, pour indiquer qu'il s'agit d'un F, nom commun.

« cada ». Exemple : *Cada um sabe o que faz*. (Chacun sait ce qu'il fait).

Ainsi, il faut prévoir une règle pour indiquer cette particularité. On peut le faire, en indiquant dans la variable RG, une valeur comme NSD.

RG =
{ VSV, NSP, PRE, NSD }NSD indique à l'analyseur que l'unité lexicale 'um' est un nom s'il est précédé par un prédéterminant « cada » ou « qualquer ».

Un autre mot, classé couramment dans la catégorie de pronoms indéfinis dans la grammaire portugaise, « qualquer » (quelconque) peut jouer le rôle d'adjectif lorsqu'il est postposé à un nom, en ce cas, il donne un sens péjoratif au nom. Exemple :

1. …un Pestana qualquer acha-se com o direito de ser deputadoJosé Lins do Rêgo. O moleque Ricardo. 5ème Edition. Rio de Janeiro :José Olympio, 1956. p. 239.. (…un Pestana quelconque se trouve avec le droit d'être député.)
2. Já não era uma Judite qualquer, era a Judite do AntunesAlmada Negreiros. Nome de Guerra. Lisboa : Verbo, 1972. p. 86.. (Elle n'était déjà pas une Judite quelconque, elle était la Judite de l'Antunes.)

Pour distinguer le mot *qualquer* qui joue le rôle d'adjectif de ce qui joue le rôle de déterminant nous allons créer une règle, c'est-à-dire une valeur qui doit être attribuée à la variable RG, pour indiquer quand il joue le rôle d'adjectif et quand il ne le joue pas.

RG =
{ VSV, NSP, PRE, NSD, ANA } où :ANA - lorsque l'unité porte cette valeur dans la variable RG, elle joue le rôle d'adjectif si elle vient après un nom.

Pour distinguer les noms communs, des noms propres et des noms « pronoms », Il faut, donc, créer une variable. À l'image du Français, on va adopter la variable NN, type de nom. Le nom propre est déjà un syntagme nominal, car il désigne un objet du monde réel.

NN =
{PRP, COM, PRO} où :PRP pour indiquer que le nom est un nom propre ;COM pour indiquer que le nom est commun ;PRO pour indiquer que le nome est du type pronom.

En portugais, il y a des noms propres qui exigent l'utilisation d'un article comme déterminants et d'autres qui le refusent. Cela arrive principalement aux noms de ville. Selon Celso CUNHA & Lindley CINTRA on utilise l'article devant un nom propre lorsqu'il était à son origine un nom commun.

Exemple :o Porto (le port), o Rio de Janeiro (la rivière), o Recife (le récif), etc.

Unité lexicale	Caté-gorie	Sous-catégories						
		NOM	PRO	GRA	NOMB	ANIM	PERS	REG
uma	F	NOM	PRO	GRA	SIN	ANI	PE1	
a	F	NOM	PRO	GRA	SIN	ANI	PE3	
da	F	NOM	PRO	MAS	SIN	ANI	PE3	-
da	F	NOM	PRO	FEM	SIN	ANI	PE3	-
nos	F	NOM	PRO	GRA	PLU	ANI	PE4	-
vos	F	NOM	PRO	GRA	PLU	ANI	PE5	-
elas	F	NOM	PRO	MAS	PLU	ANI	PE3	
elas	F	NOM	PRO	FEM	PLU	ANI	PE3	
alguem	F	NOM	PRO	GRA	SIN	ANI	PE3	-
ninguém	F	NOM	PRO	GRA	SIN	ANI	PE3	-
nada	F	NOM	PRO	GRA	SIN	ANIM	PE3	-
qualquer	F	NOM	PRO	GRA	SIN	ANIM	PE3	ANA
	D			GRA	SIN	ANIM		POSS
aquilo	F	NOM	PRO	GRA	SIN	ANIM	PE3	-
isso	F	NOM	PRO	GRA	SIN	ANIM	PE3	-
isto	F	NOM	PRO	GRA	SIN	ANIM	PE3	-
uma	F	NOM	PRO	FEM	SIN	ANIM	PE3	ANA
	D			FEM	SIN	ANIM		POSS
um	F	NOM	PRO	MAS	SIN	ANIM	PE3	ANA

Dans la figure 9.4 nous présentons une analyse complète de quelques noms du type PRO.

3.1.2.2 Rection nominale

De la même façon que les verbes, les unités de la catégorie F peuvent aussi demander des compléments, notamment ceux dérivés des verbes. L'exemple montré à cause des problèmes de structuration des SN dans le chapitre 5 section 3 (*aceitação da informação estratégica na definição do futuro da empresa* — l'acceptation de l'information stratégique dans la définition de l'avenir de l'entreprise), a été présenté comme étant de double rection. En fait, du SN originel sortent deux branches de SN indépendantes : *a informação estratégica* (l'information stratégique) et *a definição do futuro da empresa* (la définition de l'avenir de l'entreprise).

On peut supposer que le système d'identification et d'extraction automatique des syntagmes nominaux ne tient pas en compte de ce problème. Ainsi, les SN extraits

seraient :

- *a informação estratégica na defnição do futuro da empresa* (l'information stratégique dans la définition de l'avenir de l'entreprise)
- *a definição do futuro da empresa* (la définition de l'avenir de l'entreprise)
- *o futuro da empresa* (l'avenir de l'entreprise)
- *a empresa* (l'entreprise)

Dans la recherche d'information, le problème que cette démarche entraîne est de n'arriver au SN de plus haut niveau que par le centre de SN *empresa* (entreprise). Tandis qu'en prenant en compte la double rection, dans la procédure d'extraction des SN, on peut arriver au SN de plus haut niveau, soit à partir du centre de SN *informações* (informations), soit à partir du centre de SN *empresa* (entreprise). Plus que d'extraire les bons SN, c'est aussi de permettre aux usagers de trouver les documents ou les informations dont ils ont besoin.

Il faut donc utiliser la même variable que nous avons établi par les verbes (NC), pour indiquer la quantité de compléments maximaux qu'une unité de cette catégorie peut demander, en indiquant aussi les combinaisons de prépositions dans la table PREP.

Les mêmes observations faites pour les rections verbales, en ce qui concerne la suffisance de cette variable pour résoudre ce problème, sont aussi valables pour les unités de cette catégorie.

3.1.3 Prédéterminants

Catégorie représentée par le symbole « D ». Selon Alain BERRENDONNER les prédéterminants ont la propriété distributionnelle suivant :

« la propriété distributionnelle qui définit l'appartenance à cette catégorie est la possibilité pour une forme d'être employée <u>seule</u> à gauche d'un nom commun, dans un syntagme nominal ayant fonction de sujet. Soit le contexte /T__ NOM V/, où T figure une ponctuation forte. » [91] .

Jean Paul METZGER [92] remarque qu'un prédéterminant correspond à peu près, à ce qui est habituellement appelé « article ». En portugais la définition d'un prédéterminant ressemble à celle du français. Or, selon ce qui nous avons vu dans le chapitre 7, en portugais il n'y a pas d'article partitif, donc on n'a pas les problèmes posés par l'unité « des » discuté par Jean Paul METZGER. En revanche, on a le problème de l'absence des articles devant les syntagmes nominaux, objet de l'étude mené dans le chapitre 7.

Appartiennent à cette catégorie les mots suivants :

[91] ***Alain BERRENDONNER. Grammaire pour une analyseur :aspects morphologiques. Les Cahiers du Criss. Grenoble :Centre de Recherche en Informatique appliquée aux Sciences Sociales. Université des Sciences Sociales de Grenoble, Novembre, 1990. p. 25.***

[92] Jean-Paul METZGER. *Syntagmes Nominaux et Information Textuelle : reconnaissance automatique et représentation.* Thèse de Doctorat d'Etat en Sciences. Lyon : Université Claude Bernard – Lyon 1, 5 octobre 1988. p.78.

D

{a | as | o | os | dois | três | quatro | cinco | … | mil | essa | essas | esse | esses | esta | estas | este | estes | aquela | aquelas | aquele | aqueles | mesma | mesmas | mesmo | mesmos | tal | semelhante | meu | meus | teu | teus | tua | tuas | seu | sua | nosso | nossos | vosso | vossa | vossos | vossas | seus | suas | um | uma | uns | umas | alguma | algumas | algum | alguns | nenhum | nenhuma | toda | todas | todo | todos | cada | qualquer | certa | certas | certo | certos | outra | outras | outro | outros | muita | muitas | muito | muitos | pouca | poucas | pouco | poucos }

Il faut remarquer, qu'en portugais, il y a des cas où le prédéterminant est placé après le nom. Ce sont des cas des prédéterminants *algum, alguma* lesquels indiquent avoir, lorsqu'ils sont postposés au nom, la même valeur que *nenhum* (aucun), ou *nenhuma* (aucune). Exemple :

1. Não escreveu, que eu saiba, livro algumAugusto Frederico Schmidt. O galo branco : páginas de memórias. Rio de Janeiro : José Olympio, 1957. p. 71-72. . (Il n'a écrit, qu'on le sache, aucun livre)
2. …e não tinha pressa alguma de chegar em casaFerreira de Castro. Obra completa. Rio de Janeiro : Aguilar, 1958-1961. 3 v. p. 694.. (… et il n'a eu aucune hâte d'arriver chez lui).

Or, bien que Celso CUNHA & Lindley CINTRA donne à cette unité le rôle d'un déterminant, il me semble que le rôle le plus cohérent à attribuer à ce mot, dans ce contexte, est celui d'un adjectif. Pour cela, nous prenons la décision de mettre cette unité dans la catégorie F aussi, sous la catégorie ADJ, en adoptant une valeur dans la variable RG, pour indiquer à l'analyseur que ces unités joueront le rôle d'adjectif lorsqu'elle suivie un nom. La valeur ANA signale déjà cette particularité pour quelques unités. Elle peut donc être utilisée dans ce cas aussi.

RG =

{ VSV, NSP, PRE, NSD, ANA, PPN, TOD }

PPN = cette valeur indique que l'unité que la contienne, lorsqu'elle apparaît en précédant une unité F, NOM, elle jouera le rôle d'un prédéterminant. Cela pour lever l'ambiguïté des unités "o, a, os, as" qui peuvent jouer le rôle soit de prédéterminants soit le rôle d'une particule préverbale. Ainsi la caractéristique qui les désigne un rôle de prédéterminants est traduite par cette règle.

TOD est une valeur que signale à l'analyser que après les unités TODA, TODO, TODAS, TODOS, peut apparaître un déterminant défini (a, o, as, os) ou aussi un déterminant indéfini (um, uma).

Un autre aspect à repérer concerne la forme du prédéterminant, il faut savoir s'il s'agit d'un prédéterminant quantitatif (ou numéraux) ou d'autres types.

NU =

{NUM, NNU} où :NUM désigne les prédéterminants admis dans un des contextes

suivants :[___ (dos | das) (F,NOM)]SN ou [____ (F,NOM)]SNExemples : mil revistas científicas (mille revues scientifiques), qualquer (quelconque), etc.NNU désigne les autres prédéterminants.

Une autre variable important pour la reconnaissance des syntagmes nominaux est celle qui désigne le type de détermination.

TD =

{DEF, IND}DEF – prédéterminant défini (o, os, a, as, este, esta, estas, essa, essas, etc.)IND – pour désigner les prédéterminants non définis (um, algum, alguma, qualquer, etc.)

3.1.4 Particules préverbales

Catégorie représentée par le caractère Y. Font partie de cette catégorie toutes les unités qui peuvent se placer dans uns des quatre contextes suivants :

1. X __ [não] V X (dans le cas de pronoms personnels dans le rôle de sujet de la phrase : eu, tu ele, ela, nós, vós, eles elas) ;
2. X [não] __ V X (dans le cas de pronoms personnels qui jouent le rôle d'objet direct de la phrase : o, a, os, as. Et les pronoms obliquos clitique jouent le rôle d'objet indirect : me, te, nos, vous, lhe, lhes) Exemple : Eu o vi. (Je l'ai vu)
3. X V-__ X (dans le cas de pronoms personnels qui jouent le rôle d'objet direct : o, a, os, as, lhe, lhes) Exemple : Quero vê-lo (je veux le voir)
4. X V-__- (terminaison do futuro do pretérito) Exemple : vendê-lo-ia (Je le vendrais)

Tout d'abord, le X représente un terme quelconque, il peut même ne pas exister. En ce qui concerne le contexte (c), quand la forme verbale finit en -r, -s, -z on supprime ce dernier caractère et le pronom clitique assume la forme *lo, la, los, las* respectivement à (o, a, os, as). Ces mêmes pronoms pourront prendre la forme *no, na, nos, nas* respectivement lorsque des formes verbales se terminent par une double voyelle nasale comme ão, õe, em, am (ão).

Exemples :

fazer (faire) => *fazê-* ***lo*** (le faire);

encontramo- ***lo*** *em casa* (nous le rencontrons chez nous.)

Põe-na (met la)

Tem-nos (il nos tient)

Le contexte (d) est moins courant, il apparaît davantage dans les textes littéraires, et dans les poèmes, mais peu semble-t-il dans les textes techniques. Cependant, il faut le prévoir.

En ce qui concerne les pronoms de la forme *la, lo, las, los, na, no, nas, nos*, même s'il s'agit d'une forme qu'équivalent à *a, o, as, os* il ne nous semble pas nécessaire de faire un pré-traitement car ce n'est pas une contraction, mais une ajoute de caractère.

Nous allons garder chacune de ces formes dans la base de données LEXIQUE.

Les unités *o, a, os, as* sont homonymes des articles *o, a, os, as*. Il faut créer une valeur pour inclure dans la variable RG pour indiquer que ces formes sont de la catégorie Y lorsqu'elles précédent une forme verbale finie.

RG =
{ VSV, NSP, PRE, NSD, ANA, PPN, TOD, PVF } où

PVF = indique que la forme qui précède un verbe fini est une unité de la catégorie Y. Cependant, si l'unité prise est la première unité de la phrase et précède un verbe, alors là nous avons plutôt un D qu'un Y.

Il faut tenir compte aussi d'une autre particule. Il s'agit de la particule « se ». Cette particule aussi classé comme un pronom, peut jouer le rôle d'objet direct, d'objet indirect ou de sujet. De même que les autres éléments de cette catégorie cette particule est toujours à côté du verbe, soit avant, soit après. Il est donc aussi une particule préverbale.

Ainsi, sont dans la catégorie Y les formes suivantes :

Y
{eu | tu | ele | ela | nós | vós | eles | elas | o | a | os | as | no | na | nos | nas | lo | la | las | los | lhe | lhes | se }

Il est facile de voir que les éléments de cette catégorie sont des éléments anaphoriques. C'est-à-dire, ils ont une source d'anaphore. Or, la résolution de la source de ces éléments n'est pas importante pour l'extraction des syntagmes nominaux, dans le contexte de l'indexation automatique car ils ne vont pas générer des nouvelles formes de SN, mais simplement répéter leur forme originale. Cela a été déjà sûrement indexé. Ainsi, pour la recherche d'information la résolution de ce problème n'aide pas beaucoup à améliorer la précision d'un résultat d'une demande d'information. Cependant, il fallait les résoudre dans le cas des études de cooccurrence des syntagmes nominaux, dans les cas de l'analyse de contenu puisque le comptage ou non d'occurrences de SN peuvent avoir une influence importante dans le résultat de l'analyse.

Le problème des anaphores est connu, nous n'allons cependant pas nous y approfondir.

3.1.5 Prépositions

Catégorie représentée par le symbole « P ». Il n'est pas évident d'établir un contexte pour la préposition. La préposition met en rapport deux termes. Mais cela est une définition très large. Elle manque des précisions, car entre deux termes on peut avoir d'autres catégories que celle de la P. Ainsi, il nous semble très difficile de trouver un contexte particulier pour caractériser une préposition. De toute façon, on a la tentative d'Alain BERRENDONNER avec le contexte :

(même) __ SN, X [93] .

De la même manière, mais de façon plus spécifique Jean Paul METZGER suggère le contexte suivant :

(même) __ D F(NOM), …(en tête de phrase) [94] .

Il nous semble que ces deux contextes suggérés soient valables aussi pour le portugais, on peut les adapter de la façon suivante :

(proprio) __D F(NOM), … (en tête de phrase) ou

(proprio) __SN, X

En fait, il s'agit d'une catégorie très large. Ce qui met en opposition les grammairiens qui la classe en catégorie ouverte et ceux qui la classe en une catégorie fermée. Par rapport aux prépositions simples, à notre avis elles peuvent être en catégorie fermée. Le problème est avec les prépositions composées, en effet, les locutions prépositionnelles. Ce sont les locutions prépositionnelles qui font que cette catégorie soit considérée même comme étant ouverte. Une idée de cette ouverture est donné par Celso CUNHA & Lindsey CINTRA lorsqu'ils définent les locutions prépositionnelles comme étant des mots ***« constitués de deux ou plus unités lexicales, la dernière étant une préposition simple (généralement de) »*** [95] .

Etant donné qu'il n'y a pas une caractéristique des mots qui puissent constituer une préposition ou mieux, une locution prépositionnelle, et bien qu'elles soient nombreuses nous les enregistrerons dans la base de données LEXIQUE, tant les prépositions simples ou essentielles que les locutions prépositionnelles. Cela doit alléger la syntaxe d'extraction des syntagmes nominaux.

Inventaire des prépositions les plus couramment utilisés :

P □

{a | de | para | por | per | com | sem | em | ante | após | entre | contra | sob | sobre | até | desde | perante | trás | abaixo de | acerca de | acima de | a despeito de | adiante de | a fim de | além de | antes de | ao lado de | ao redor de |a par de | apesar de | a respeito de | atrás de | através de | de acordo com |debaixo de | de cima de | defronte de | dentro de | depois de | diante de |em baixo de | em cima de | em frente de | em lugar de | em redor de |em torno de | em vez de | graças a | junto a | junto de | para baixo de |para cima de | para com | perto de | por baixo de | por causa de | por cima de |por detrás de | por diante de | por entre | por trás de | em relação a…}

En regardant les locutions prépositionnelles, on se rend compte du besoin d'une recherche sur la syntaxe des locutions prépositionnelles de façon à rendre possible leur

[93] Alain BERRENDONNER. *Grammaire pour une analyseur :aspects morphologiques.* Les Cahiers du Criss. Grenoble :Centre de Recherche en Informatique appliquée aux Sciences Sociales. Université des Sciences Sociales de Grenoble, Novembre, 1990. p. 25.

[94] Jean-Paul METZGER. *Syntagmes nominaux et information textuelle :reconnaissance automatique et représentation.* Thèse de doctorat d'Etat en Sciences. Lyon : Université Claude Bernard - Lyon 1, 1988. p. 80.

[95] Celso CUNHA & Lindsey CINTRA. *Nova gramática do português contemporâneo.* Lisboa : Edições João Sa da Costa, 1984. p. 551.

analyse sans les garder dans la base de données LEXIQUE. Ce qui donnera une vue plus générale à l'analyseur et rendra indépendant d'un lexique.

3.1.6 Conjonctions de coordination

Cette catégorie est représentée par le symbole « C ». Nous allons suivre la proposition fait pour le français par Alain BERRENDONNER. Il ne considère comme conjonctions de coordination que les mots qui sont capables de connecter non seulement deux phrases, mais aussi deux termes (et plus précisément, des syntagmes nominaux). Selon lui, « *le critère d'identification d'une conjonction de coordination est l'aptitude à apparaître de manière indéfiniment récurrent, devant les n termes nominaux d'une énumération*. » [96] . En portugais on trouve des mots qui apparaissent dans ce contexte. Ils peuvent être inventoriés comme étant :

C □
{e | ou | ora | quer | seja | nem}

Et de la même façon que pour le français, ***« tous connecteurs qui ne peut lier que deux phrases successives, ou bien deux expressions prédicatives, seront traités comme un adverbe anaphorique »*** [97] . Exemples : mas, porém, porque, pois, como [=porque], pois que etc.

3.1.7 Conjonctions de subordination

Cette catégorie est représentée par le symbole « Q ». Les travaux menés par Alain BERRENDONNER et Jean-Paul METZGER, ont réduit les unités de cette catégorie, pour l'essentiel, à {que, de, si} et le second a ajouté {à}. La régularisation de toutes les formes de pronoms relatifs et de conjonctions de subordination les a réduits à {que}. Nous devrions faire la même chose pour la langue portugaise. Cependant, cette catégorie y est très vaste.

Cependant, dans notre corpus nous n'avons pas extrait une quantité important de SN avec ces unités. Nous en avons trouvé quelques-uns uns avec la conjonction « *como* » (comme). Une étude minutieuse de ces unités permettrait de mieux connaître leurs propriétés distributionnelles.

En ce que concerne les pronoms relatifs, parmi les SN extraits, il n'y avait que 287 SN avec des pronoms relatifs, ce qui représente 4,8% du total de 5982 SN. Ce n'est pas une proportion trop importante. Il est quand même souhaitable de les repérer, car ce sont des SN qui peuvent appartenir à ceux de plus haut niveau, donc important pour le raffinement d'une recherche d'information.

Conscient de l'importance de la régularisation des unités qui appartiennent aux

[96] Alain BERRENDONNER. *Grammaire pour une analyseur :aspects morphologiques*. Les Cahiers du Criss. Grenoble :Centre de Recherche en Informatique appliquée aux Sciences Sociales. Université des Sciences Sociales de Grenoble, Novembre, 1990. p. 26.

[97] Ibidem p. 27.

pronoms relatifs et aux conjonctions de subordination, nous allons mettre les pronoms relatifs et les conjonctions de subordination tels qu'ils sont, provisoirement dans cette catégorie. Ceci pour plusieurs raisons : a) les unités qui appartiennent à ces deux catégories, dans la langue portugaise, sont nombreuses ; b) il faut bien connaître leurs propriétés distributionnelles, donc une étude plus minutieuse ; c) la régularisation des unités qui appartiennent à cette catégorie n'est pas évidente ; d) les conjonctions de subordination ne font pas partie de la syntaxe de syntagme nominal développée dans ce travail.

Cette étude peut être le sujet d'une autre recherche à la suite de ce travail. Cette étude sera terminée, il ne restera qu'à faire la mise à jour du champ RG (règles) de chaque unité correspondante à cette catégorie dans la base de données LEXIQUE, en mettant la valeur de la règle de régularisation correspondante. Cependant, il faut spécifier aussi quelles sont les unités résultant de cette régularisation.

Q

{que | quanto | onde | cujo | se | caso | como | assim como | porque | para que | contanto que | salvo se | sem que [=se não] | conforme | segundo |consoante | quanto mais | à proporção que | que nem | bem como |dado que | desde que | a menos que | a não ser que | quando | antes que | cada vez que | de maneira que | de sorte que | de forma que, etc…}

Il est possible de voir que beaucoup de conjonctions de subordination peuvent être réduites à la forme / que / plus une autre unité lexicale (comme des prépositions ou des adverbes, etc.). De la même façon, il y a des pronoms relatifs équivalents aux pronoms relatifs français, comme : *que* (que), *quem* (qui), *cujo* (dont), *o qual* (lequel). Il est vrai qu'on pourrait faire une sorte d'équivalence entre ces unités, mais il nous semble beaucoup plus prudent de faire une analyse plus minutieuse de chaque unité, autant pour les pronoms relatifs que pour les conjonctions de subordination.

3.1.8 Adverbes

Cette catégorie est représentée par le symbole W. On y englobe les adverbes et aussi tous les mots qui ne sont pas compatibles avec d'autres catégories. Pour cela, on dit qu'il s'agit d'une catégorie poubelle. Mais dans une telle catégorie, les unités peuvent encore être classées dans quelques sous-catégories.

Les adverbes de quantités, ce sont des unités qui peuvent faire partie d'un déterminant : muito, pouco, mais, menos, etc.

Les adverbes d'intensité, ce sont des unités qui modifient les adjectifs : assaz, bastante, bem, demais, mais, menos, muito, pouco, quanto, quão, quase, tanto, tão, etc.

Les adverbes de temps, aspects et mode susceptibles d'occuper des positions syntaxiques originales : possivelmente, sinceramente, afirmativamente, provavelmente, sim, quiçá, talvez, depois, agora, ainda, hoje, antes, breve, cedo, tarde, já, etc. Et aussi, d'autres terminés par le suffixe -mente.

Les adverbes anaphoriques : ali, assim, lá, abaixo, acima, adiante, aí, aqui, etc.

Les adverbes de négation : não.

En langue portugaise, la négation est marquée par une seule particule alors qu'en français il en faut fréquemment deux.

W □
{ muito | pouco | sim | quiçá | talvez | depois | agora | ainda | hoje | antes | breve | cedo | tarde | já | ali | assim | lá | abaixo | acima | adiante | aí | aqui | não, etc}

La variable type d'adverbes TW aura les valeurs suivantes :

TW =
{QUA, AAJ, PRO, TAM, NEG} où :QUA indique que l'unité en question est un adverbe de quantité ;AAJ indique que l'unité en question est un adverbe qui modifie un adjectif ;PRO pour indiquer les adverbes anaphoriques ;TAM pour indiquer les adverbes de temps, aspects et mode.NEG pour indiquer que l'unité en question est un adverbe de négation. (não)

D'une manière distincte de celle adoptée par Alain BERRENDONNER, nous ne créons pas une catégorie spécifique pour l'adverbe de négation 'não'.

Les mots « muito » et « pouco » sont aussi dans la liste des prédéterminants. Ils jouent le rôle d'adverbe, lorsqu'ils précédent des adjectifs, des verbes ou des adverbes. Il faut donc créer une règle dans la variable RG pour exprimer cette particularité.

RG =
{ VSV, NSP, PRE, NSD, ANA, PPN, TOD, PVF, WAV }WAV = indique au module SIDUL que l'unité prise sera dans la catégorie W si elle précède un verbe, adjectif ou un élément de la catégorie W.

3.1.9 Ponctuation

Cette catégorie est représentée par le symbole T. Font partie de cette catégorie les signes de ponctuation utilisées normalement en langue portugaise écrite.

T □
{. | ! | ? | , | : | ; | - | ' | « | » | / | (|) |…}

Ainsi que pour les autres unités il a fallu créer des variables pour mieux caractériser ces différents signes car chacun d'eux a sa spécificité et son rôle dans les textes. On appellera cette variable de VP et les valeurs qu'on peut lui attribuer sont :

VP =
{PSP, PSM, PDP, PSU} où :PSP indique à l'analyseur que l'unité reconnue est une forme de ponctuation marquant une frontière de phrase {. | ? | ! } ; PSM indique à l'analyseur que l'unité reconnue est une forme de ponctuation marquant la séparation de membres d'une phrase {, | ; | :} ;PDP indique à l'analyseur que l'unité reconnue est une partie d'une double ponctuation. C'est-à-dire, une ouverture ou une fermeture, soit de guillemets, soit de parenthèses, soit de tirets

(sachant que la fermeture d'un membre de phrase initié par un tiret peut être un point final) ; PSU indique la fin d'une énumération (point de suspension /…/) et a le rôle d'un point marquant la frontière d'une phrase. En ce cas, il nous semble important qu'elle appartienne au SN, simplement pour laisser la marque d'une probable séquence de l'énumération.

En ce qui concerne le caractère « / », il peut jouer un rôle de ponctuation ou d'une conjonction de coordination {ou}. Par exemple : a) Dans le cas de « *S/A* » ce caractère joue le rôle d'un point ; b) Dans le cas de « *arquivos / bibliotecas / centros de informação* » (« fichiers / bibliothèque / centre d'information ») ce caractère joue le rôle d'une conjonction {ou} ; c) On trouve des cas ambigus aussi : « *e/ou* » (« et/ou ») comme dans « *informação científica e/ou tecnológica* » (« information scientifique et/ou technologique »). Il peut s'agir d'une information scientifique et technologique à la fois ou d'une information scientifique ou d'une information technologique alternativement. Il faut laisser l'ensemble d'unités « e/ou » comme il est sans changer ce caractère car si on fait un échange, on risque de se tromper en extrayant des faux syntagmes nominaux.

En ce qui concerne le trait-d'union selon la discussion au début de ce chapitre, il fait partie de mots composés et il est aussi utilisé pour lier un verbe fini à un pronom clitique.

3.1.10 Nombres

Nous allons représenter cette catégorie par le caractère E. Bien que nous ne traitions que des textes, dans ce travail, il faut prévoir que dans un texte on peut trouver des nombres, soit pour indiquer des dates, soit pour désigner des nombres comme la quantité d'octets d'une mémoire d'ordinateur, la quantité d'espace en disque, ou le coût d'investissement dans un secteur, etc. Ainsi, il faut prendre en compte ces types d'unités. C'est une catégorie particulière qui peut jouer plusieurs rôles dans un syntagme nominal, soit le rôle de prédéterminant (15 livres d'amour), soit d'un nom (la décennie de 50). Un nombre peut être aussi un syntagme nominal, lorsqu'il représente une date (1997, 1998).

Il est nécessaire de repérer ce genre de donnée puisqu'il peut entrer dans la syntaxe de syntagmes nominaux. La question que se pose est : il faut enregistrer ce genre d'unité dans la base de données LEXIQUE ? Or, ces données sont très variées et peuvent provoquer une croissance importante de la base de données. De plus, elles sont de faible utilisation si on les compare avec les unités lexicales (mots). Il nous semble que la meilleure solution serait le passage de ces données identifiées et caractérisées directement au module de reconnaissance et d'extraction de SN. Au module de segmentation SIDUL revient la tâche d'identifier et de caractériser les données numériques. Nous voyons là l'importance de créer cette catégorie. Il faut fournir encore une autre information au module REIS de reconnaissance et d'extraction de SN, pour indiquer s'il s'agit d'un chiffre entier ou décimal. Pour cela, nous allons créer la variable TC, type de nombre dont les valeurs sont :

TC =
{INT, DEC} où :INT indique que le nombre est entier ;DEC indique que le nombre est décimal, c'est-à-dire le nombre est du type 9.999,99.

Le module SIDUL doit indiquer aussi, dans la variable RG, une valeur PGN, pour que le module REIS puisse vérifier à quoi se rapporte le nombre. En fait cette valeur indique que le nombre sera un prédéterminant s'il apparaît à gauche d'une unité qui appartient à la catégorie F, sous-catégorie NOM, sinon il peut être à la place d'une date ou même d'un nom.

3.2 Eléments anaphoriques

Tout d'abord, on trouve dans les textes en langue portugaise autant d'éléments anaphoriques que dans les textes en langue française. Mais, qu'est-ce qui une anaphore anaphore ? Isabelle VIDALENC-SABOURIN donne, dans sa thèse, une définition basée sur celle donnée par O. DUCROT, dans un article sur les « Relations sémantiques entre phrases », page 358, du dictionnaire encyclopédique des sciences du langage.

Selon Isabelle VIDALENC-SABURIN : ***« L'anaphore est une relation discursive entre une unité et une autre antécédente basée sur le fait que la première unité ne peut à elle seule avoir une interprétation. Il n'y a anaphore entre deux unités textuelles que si l'on peut dire : l'unité antécédente apporte un élément qui permet à l'unité anaphorique d'être interprétée. »*** [98] . De la même façon qu'en français, cette figure existe aussi en langue portugaise, par le biais des particules préverbales (Y), des pronoms possessifs, parmi d'autres unités. Il s'agit d'un phénomène fréquent dans les textes. Pourtant, la question que nous voulons poser est : faut-il faire des efforts pour résoudre tous les cas d'anaphores rencontrés dans les textes ? Si les unités anaphoriques font référence à une unité antécédente, est-il important de les repérer ?

Il nous semble qu'on peut partager les anaphores en deux situations différentes : soit elle fait simplement référence à une unité précédente sans former un nouveau syntagme nominal, soit le fait en formant un nouveau SN. Dans les deux situations, nous supposons que l'unité précédente est un syntagme nominal. La première situation a comme conséquence un repérage doublé du syntagme nominal, il n'apporte donc rien de nouveau à la procédure d'indexation automatique ni à la procédure de recherche d'information, car le syntagme nominal a été déjà repéré. Un autre côté, dans la deuxième situation, l'insertion de l'unité précédente contribue à la formation d'un nouveau syntagme nominal, donc cette opération est intéressante pour l'indexation automatique et aussi pour la recherche d'information. Elle peut aider à la procédure de raffinement d'une recherche d'information, étant donné que le nouveau syntagme nominal sera, en conséquence, d'un niveau supérieur. Ce qui représente une possibilité d'améliorer le résultat d'une recherche d'information.

Bien que le premier contexte n'apporte pas de nouveau syntagme nominal, il peut quand même être utile à l'analyse de contenu, car le comptage des syntagmes nominaux peut être important dans la procédure d'un calcul d'ordre de pertinence ou d'importance d'un syntagme nominal par rapport à d'autres syntagmes nominaux. Une réflexion purement basée sur le corpus que nous avons travaillé dans le cadre du DEA, nous pouvons dire, grosso modo, que les anaphores dues aux prédéterminants possessifs sont

[98] Isabelle VIDALENC-SABOURIN. *Traitement automatique des anaphores en français : étude linguistique préalable.* Thèse de doctorat en Sciences de l'Information et Communication. Lyon : Université Lumière – Lyon 2, janvier 1989. p. 37.

en grande partie responsables pour la formation de nouveaux syntagmes nominaux. Tandis que les anaphores dues aux formes préverbales (Y) ne contribuent pas à la formation de nouveaux syntagmes nominaux.

Il faut rappeler que la formation d'un nouveau syntagme nominal due à la résolution d'un élément anaphorique doit être faite de manière attentive car on peut arriver à des résultats curieux, selon les commentaires faits dans le chapitre 3 [99] . D'ailleurs Isabelle VIDALENC-SABOURIN a fait aussi des remarques là-dessus, en donnant des exemples [100] à la page 36 de sa thèse.

Ainsi, nous croyons qu'il faut bien définir les critères de résolutions des sources d'anaphores, tenant toujours compte du but du travail.

Nous avons décidé de ne pas travailler sur les problèmes d'anaphores pour l'instant, étant donné la complexité et extension du problème. Pourtant, la réalisation d'une étude approfondie sur ce problème envisageant l'indexation automatique et la recherche d'information est essentielle. Les thèses de Isabelle VIDALENC-SABOURIN et Valérie LARROCHE-BOUTET nous semblent importantes comme référence à une étude des anaphores en langue portugaise, étant donné les ressemblances de la langue portugaise avec la langue française.

3.3 Consolidation des catégories et variables

Afin d'organiser et de résumer ce chapitre, nous allons consolider les catégories définies et les variables établies pour caractériser les unités lexicales.

En ce qui concerne la caractérisation majeure, celle des unités lexicales, nous en avons établi 10 (dix) différentes, qui ne recouvrent pas exactement celles définis pour la langue française, du fait de l'inexistence de la catégorie G, celle de la particule de négation (ne), laquelle correspond en portugais à « *não* ». Nous obtenons les catégories suivantes :

F
Nominaux ;

V
Verbes ;

D
Prédéterminants ;

Y
Particules préverbales ;

[99] Un syntagme avec un élément anaphorique : « une catégorie des clients conscientisés sur leurs droits à des produits et à des services de haute qualité » ; lLe nouveau SN réécrit après la solution de l'élément anaphorique : « une catégorie des clients conscientisés sur les droits des clients conscientisés à des produits et à des services de haute qualité ».

[100] La phrase avec l'élément anaphorique : « Tous les concurrents espèrent qu'ils vont gagner » ; La phrase après la solution de l'élément anaphorique : « Tous les concurrents espèrent que tous les concurrents vont gagner » ; ou « J'ai rencontré des amis ; ils m'ont parlé de toi » ; « J'ai rencontré des amis ; des amis m'ont parlé de toi »

P
Prépositions ;

C
Conjonctions de coordination ;

Q
Conjonctions de subordination ;

W
Adverbes ;

E
Nombres ;

T
Ponctuations.

En ce qui concerne les variables de sous-catégorisation nous avons défini :

· Les variables de sous-catégorisation syntaxique

1. NA – Type Nominaux dont les valeurs sont : NA = {NOM, ADJ, NAN} ; NOM – pour indiquer le type Noms, substantifs ; ADJ – pour indiquer le type Adjectif ; NAN – non marqués.
2. PG – Participe / Gérondif dont les valeurs sont : PG = {PAR, GER} ; PAR – pour indiquer que le verbe est au Participe ; GER – pour indiquer que le verbe est au Gérondif.
3. VB – Formes verbales dont les valeurs sont : VB = {INF, FIN} ; INF – pour indiquer que le verbe est à l'Infinitif ; FIN – pour indiquer que le verbe est fléchi ou fini.
4. VX – Type de verbe dont les valeurs sont : VX = {AUX, ORD} ; AUX – pour indiquer que le verbe est un verbe auxiliaire [ter(avoir), haver(avoir), ser(être) et estar(être)] ; ORD – pour indiquer tous les autres verbes.

· Les variables de sous-catégorisation flexionnelles

1. GR – Flexion en Genre dont les valeurs sont : GR = {MAS, FEM, GRN} MAS pour indiquer que l'unité est au masculin ; FEM pour indiquer que l'unité lexicale est au féminin ; GNR pour indiquer que l'unité lexicale est non marquée.
2. NB – Flexion en Nombre dont les valeurs sont : NB = {PLU, SIN, NBN} PLU pour indiquer que l'unité est au pluriel ; SIN pour indiquer que l'unité est au singulier ; NBN pour indiquer que l'unité est non marquée.
3. PE – Flexion en Personne dons les valeurs sont : PE = {PE1, PE2, PE3, PE4, PE5} PE1 indique la première personne du singulier, cas de eu (je) ; PE2 indique la deuxième personne du singulier, tu (tu) ; PE3 indique la troisième personne, cas de ele, ela, eles, elas (il, elle, ils, elles) ; PE4 indique la première personne du pluriel, cas de nós (nous) ; PE5 indique la deuxième personne du pluriel, cas de vós (vous).

· Les variables de sous-catégorisation lexicale

1. NN – Type de Noms dons les valeurs sont : NN = {PRP, COM, PRO} PRP – pour indiquer les noms propres ; COM – pour indiquer les noms communs ; PRO – pour indiquer les pro-formes nominales.

2. VN – Type de nom commun dont les valeurs sont : VN = {CON, ABS } CON – pour indiquer les noms concrets ; IMM – pour indiquer les noms abstraits.

3. NU – Type de prédéterminants dont les valeurs sont : NU = {NUM, NNU} NUM – pour indiquer les prédéterminants quantitatifs ou numéraux ; NNU – pour indiquer les autres prédéterminants.

4. TD – Type de détermination dont les valeur sont : {DEF, IND} DEF – prédéterminant défini (o, os, a, as, este, esta, estas, essas, etc.) IND – pour désigner les prédéterminants non définis (um, algum, alguma, qualquer, etc.)

5. AN – Type d'animation dont les valeurs sont : AN = {ANI, INA, ANN } ANI – pour indiquer le caractère animé ; INA – pour indiquer le caractère inanimé ; ANN – pour indiquer le caractère non marqué.

6. TC – Type de nombre dont les valeurs sont : TC = {INT, DEC} INT – pour indiquer les nombres entiers ; DEC – pour indiquer les nombres décimaux.

7. NC – Nombre de compléments dont les valeurs sont : NC = {0CO, 1CO, 2CO, 3CO, 4CO}Les valeurs indiquent le maximum de compléments qu'une unité lexicale peut demander. 0CO – indique que l'unité n'exige aucun complément ; 1CO – indique que l'unité exige un complément ; 2CO – indique que l'unité exige deux compléments ; 3CO – indique que l'unité exige trois compléments ; 4CO – indique que l'unité exige quatre compléments.

8. TA – Type d'Adjectifs dont les valeurs sont : TA = {QUA, REL} QUA indique les adjectifs de qualité ; REL indique les adjectifs de relation.

9. TW – Type d'Adverbes dont les valeurs sont : TW= {QUA, AAJ, PRO, TAM, NEG} QUA indique que l'unité en question est un adverbe de quantité ; AAJ indique que l'unité en question est un adverbe qui modifie un adjectif ; PRO pour indiquer les adverbes anaphoriques ; TAM pour indiquer les adverbes de temps, aspects et mode. NEG pour indiquer que l'unité en question est un adverbe de négation. (não).

10. VP - Type de ponctuation dont les valeurs sont : VP= {PSP, PSM, PDP, PSU} PSP indique à l'analyseur que l'unité reconnue est une sorte de ponctuation marquant une frontière de phrase {. | ? | ! } ; PSM indique à l'analyseur que l'unité reconnue est une sorte de ponctuation marquant la séparation des composants d'une phrase {, | ; | :} ; PDP indique à l'analyseur que l'unité reconnue est une partie d'une double ponctuation. C'est-à-dire, soit un guillemet d'ouverture ou de fermeture, soit une ouverture ou une fermeture de parenthèses, soit une ouverture ou fermeture de tiret (sachant que la fermeture d'un membre initié par un tiret peut être un point final) ; PSU indique la fin d'une énumération (point de suspension /…/) et a le rôle d'un point marquant la frontière d'une phrase. En ce cas, il nous semble important qu'il appartienne au SN, simplement pour laisser la marque de fin d'une probable

séquence d'énumération.

· Les variables d'utilisation générique

1. RG – Type de règles dont les valeurs sont : RG = { VSV, NSP, PRE, NSD, ANA, PPN, TOD, PVF, PGN, WAV } VSV - indique au module SIDUL que l'unité sera interprétée comme une forme verbale si elle vient après un verbe auxiliaire. NSP - indique au module SIDUL que dès qu'il y a un prédéterminant à gauche d'un verbe à l'infinitif, cette forme verbale joue alors le rôle d'un FNOM ; PRE – indique au module SIDUL qu'il faut faire un pré-traitement sur l'unité prise. Cela signifie qu'il faut faire une régularisation de l'unité en question, étant donnée qu'il s'agit d'une sorte d'amalgame. L'action de régularisation est faite à partir de la reconstitution de l'unité en ses composants originels, lesquels doivent être informés dans la table ECHANGE ; NSD – indique à l'analyseur que l'unité lexicale 'um' est un nom s'il est précédé par un prédéterminant « cada » ou « qualquer ». ANA – indique au module SIDUL que l'unité prise, doit être prise comme un adjectif si elle vient après un nom ; PPN – indique au module SIDUL que l'unité prise, sera un prédéterminant si elle précède une unité F, NOM ; TOD – signale à l'analyser que après les unités TODA, TODO, TODAS, TODOS, peut apparaître un déterminant défini (a, o, as, os) ou aussi un déterminant indéfini (um, uma) ; PVF – indique au module SIDUL que l'unité prise, sera une unité de la catégorie Y si l'unité qu'elle précède est un verbe fini. Cependant elle peut être un D (prédéterminant) si elle est la première unité de la phrase ; PGN – indique que l'unité est un nombre et qui, s'il est à gauche d'un nom il joue le rôle d'un prédéterminant. WAV - indique au module SIDUL que l'unité prise sera dans la catégorie W si elle précède un verbe, un adjectif ou un élément de la catégorie W.

4 Structure de la base de données LEXIQUE

La base LEXIQUE sera structurée selon le modèle relationnel. L'entité majeure dans cette base est l'unité lexicale (mot, mot composé, ponctuation). Cette structure, dont on a beaucoup discuté précédemment, dépend des faits qui caractérisent cette entité et qui sont :

1. Chaque unité lexicale a une catégorie, représentée par un caractère ;
2. Chaque unité lexicale a un ensemble de caractéristiques appelées variables de sous-catégorisation, soit d'ordre lexical, flexionnelle ou syntaxique ;
3. Chaque unité lexicale peut appartenir à plus d'une catégorie ;
4. Chaque unité lexicale peut avoir une règle de contrainte appelée RG ;
5. Chaque unité lexicale peut exiger un ou plusieurs compléments. Ces compléments peuvent être régis ou non (cas d'objet direct) par une préposition. Dans ce cas, il faut indiquer les prépositions susceptibles d'apparaître pour chaque complément ;
6. Certaines unités lexicales sont le résultat d'une contraction de deux autres unités

lexicales qui appartiennent à des catégories déjà existantes. En ce cas, il faut réaliser un pré-traitement pour restituer leurs unités originelles envisageant leur traitement individuel par l'analyseur. Ce pré-traitement devra être indiqué dans l'enregistrement de l'unité lexicale contractée, à travers l'indication de la valeur PRE dans la variable RG ;

Après ces faits explicités plus haut, on voit que les quatre tables suivantes sont nécessaires :

1. ULEX (code, unité_lexicale) où : code : := code numérique qui identifie chaque unité lexicale. Cet attribut est la clé primaire de cette relation ; unité_lexicale : := mot, mot composé ou ponctuation. Cet attribut doit être indexé.
2. CARACTERISTIQUES(code, catégorie, var, valeur) où : code : := contient le code d'identification de l'unité lexicale. Cet attribut et l'attribut catégorie forment la clé primaire ; catégorie : := cet attribut contient la catégorie de l'unité lexicale ; var : := représente chaque variable qui peut contenir des valeurs qui caractérisent une unité lexicale. Si une unité a plusieurs caractéristiques, c'est-à-dire plusieurs sous-catégories, il y aura autant de tuplesRappelons que tuple est un ensemble d'attributs d'une table, en ce cas formé par <code, catégorie, var, valeur> Exemple : <1,F, GR, FEM> , <1, F, NB, SIN>, <1, F, NC, 2>. que le nombre de sous-catégories. Exemple de var ou variable : GR = genre, NB = nombre, AN = animation, NC = nombre de compléments ; valeur : := la valeur correspondante à la variable explicitée dans une tuple.
3. PREP (code, coderc, prp1, prp2, prp3, prp4) où : code : := contient le code d'identification de l'unité lexicale ; coderc : := cet attribut contient un code numérique séquentiel. Cet attribut avec l'attribut code forment la clé primaire de cette relation. Cet attribut doit exister car on peut avoir plus d'une combinaison de prépositions par chaque unité lexicale ; prp1 : := contient la préposition qui doit précéder le premier complément de l'unité lexicale représenté par le code. Si le complément est un objet direct, cet attribut doit avoir la valeur blanche ; prp2 : := contient la préposition qui doit précéder le deuxième complément de l'unité lexicale représenté par le code. Cet attribut ne doit être rempli que lorsque la valeur de la variable NC dans la relation CARACTERISTIQUE est 2CO ; prp3 : := contient la préposition qui doit précéder le troisième complément de l'unité lexicale représenté par le code. Cet attribut ne doit être rempli que lorsque la valeur de la variable NC dans la relation CARACTERISTIQUE est 3CO ; prp4 : := contient la préposition qui doit précéder le troisième complément de l'unité lexicale représenté par le code. Cet attribut ne doit être rempli que lorsque la valeur de la variable NC dans la relation CARACTERISTIQUE est 4CO.
4. ECHANGE(code, code1, code2) où code : := contient le code de l'unité lexicale contractée. C'est aussi la clé primaire de cette relation ; code1 : := contient le code de la première unité lexicale participante de l'unité lexicale contractée, suivie de sa catégorie pour qu'on puisse retrouver ses caractéristiques dans la relation CARACTERISTIQUES ; code2 : := contient le code de la deuxième unité lexicale

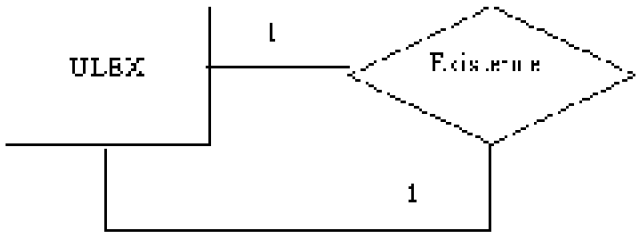
participante de l'unité lexicale contractée, suivie de sa catégorie pour qu'on puisse retrouver ses caractéristiques dans la relation CARACTERISTIQUES.

4.1 Esquisse de la structure de la base de donnée LEXIQUE

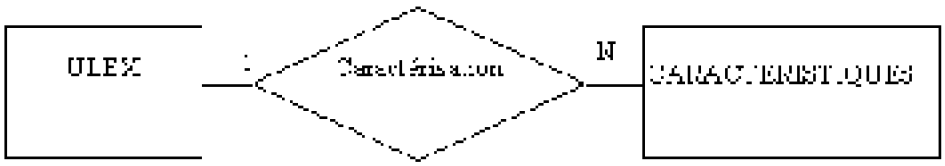
La démarche à suivre est celle de l'identification de l'unité lexicale, une fois que le module SIDUL a fait la segmentation du texte, c'est-à-dire qu'il a pris une unité lexicale, il doit accéder à la base de données pour vérifier si l'unité extraite existe ou non. C'est le rapport d'existence, on le montre dans la figure 9.5.

Les numéros (1) présentés dans la figure 9.5 indiquent qu'à chaque unité lexicale correspond 1 seul code. Une fois vérifiée l'existence d'une unité lexicale : si le mot existe, il faut chercher ses caractéristiques ; sinon il faut l'enregistrer avec toutes celles-ci.

Il faut entamer une procédure de création de l'enregistrement de l'unité si elle n'existe pas encore dans la base LEXIQUE.



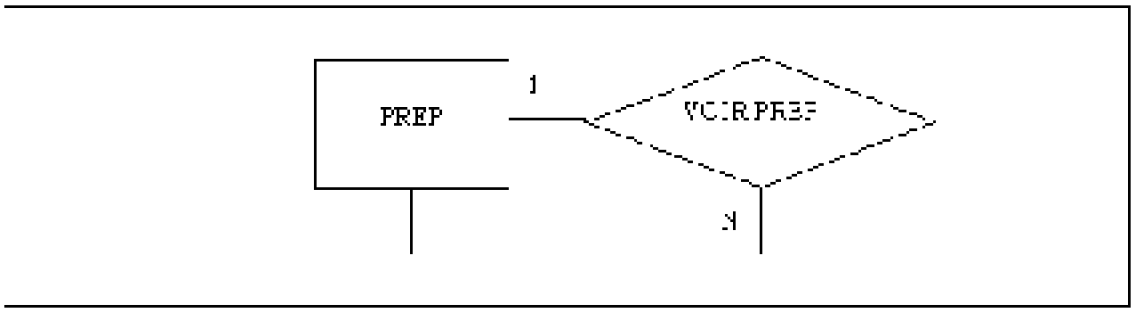
Une fois que l'unité cherchée existe et, si on a le code d'identification de cette unité, il faut un autre rapport pour trouver ses caractéristiques. Dans la figure 9.6, on montre ce rapport, appelé CARACTERISATION.



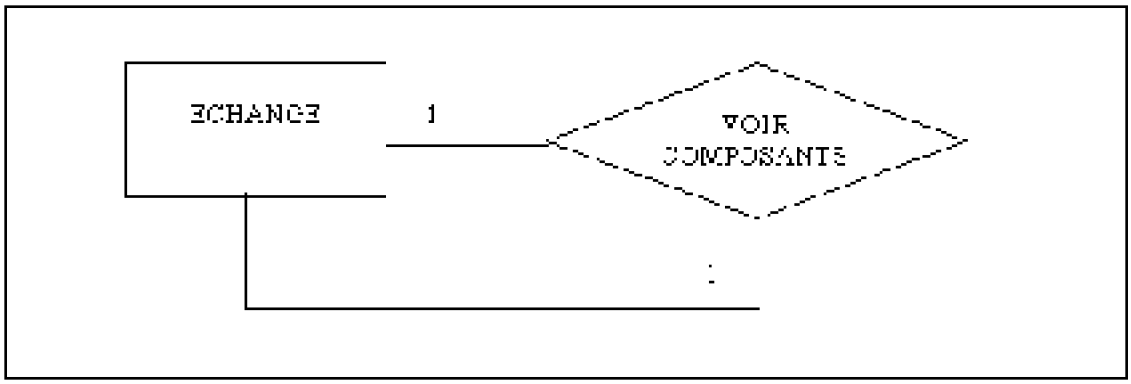
Dans la figure 9.5, le numéro 1 à coté de la relation ULEX et le caractère N à côté de la relation CARACTERISTIQUES montrent qu'avec chaque unité lexique on peut avoir plus d'un enregistrement de caractéristiques. C'est parce que des unités présentent la même forme et participent à des catégories différentes. Exemple : 'o' peut être aussi bien un prédéterminant (D) qu'une particule préverbale (Y). A partir de ce rapport l'analyseur aura toutes les variables d'une unité lexicale donnée, c'est-à-dire toutes les caractéristiques prévues par la grammaire qu'on est en train de concevoir.

Pourtant, il existe des variables que nécessitent de chercher plus d'informations. Exemple : la variable NC a une valeur parmi 1CO, 2CO, 3CO et 4CO, cela signifie qu'il

faut encore savoir quelles sont les prépositions susceptibles d'apparaître avec chaque complément. Pour trouver ces prépositions on a le rapport VOIR PREP, montré par la figure 9.7.



Cette figure montre encore qu'avec chaque unité lexicale on peut avoir plus d'une combinaison de prépositions, ce qui est indiqué par le caractère N en bas du losange.



D'autres informations peuvent être nécessaires comme les unités constituantes d'une unité lexicale contractée. Lorsqu'on trouve dans la variable RG, attribut de la relation ULEX, une valeur PRE, cela signifie que l'unité lexicale identifiée est un mot contracté et qu'il faut faire un pré-traitement, c'est-à-dire qu'il faut restituer ses composants pour pouvoir les traiter individuellement. Pour cela, on utilise le rapport VOIR COMPOSANTS, montré dans la figure 9.7.

Les numéros 1 à coté de chaque relation indique qu'à chaque unité contracté il n'y a qu'un seul ensemble de composants. A partir des données obtenues par le rapport VOIR COMPOSANTS l'analyseur peut retrouver les caractéristiques de chaque composant à travers le rapport Caractérisation.

5 Conclusion

Au lieu de faire une description exhaustive des unités lexicales de la langue portugaise, nous avons préféré faire une description plus simple mais qui permettra la reconnaissance des syntagmes nominaux et leur extraction, le but premier de ce travail. Il est vrai que nous avons laisser quelques aspects de côté comme les éléments anaphoriques, les pronoms relatifs et les conjonctions de subordination.

En ce qui concerne les éléments anaphoriques nous avons déjà expliqué les raisons

de ne pas travailler actuellement ce point, c'est à la suite de cette décision que nous n'avons pas créé de variables pour aider la résolution des problèmes des anaphores. Il nous semble, d'autre part, qu'outre les variables PE et AN, il faudrait en établir d'autres pour représenter des traits sémantiques visant à aider la résolution des anaphores. En effet, ce sont des variables à être défini par l'étude de résolution des anaphores.

Par rapport aux pronoms relatifs et aux conjonctions de subordination, ces unités ne font pas partie de la grammaire d'extraction des syntagmes nominaux. Nous envisageons de les traiter plus tard dans une étude plus approfondie. Ceci parce que la syntaxe concernant les syntagmes nominaux ayant des pronoms relatifs et des conjonctions de subordinations sont trop diversifiés. Il faut d'abord vérifier la possibilité de régulariser les unités de ces deux catégories et essayer de les réduire comme on a procédé pour le français. Cette procédure de réduction peut aider à définir une syntaxe pour ce type de syntagme nominal. Bien que les formes syntaxiques de ce type de syntagme nominal soit très disperse, le nombre total n'est pas trop élevé dans le corpus, il ne représente que 4,8% de tous les syntagmes nominaux.

D'une manière générale nous avons crée des variables qui sont capables d'aider à reconnaître des unités qui constituent un syntagme nominal. Outre les variables caractérisées dans l'axe de la syntaxe, du lexique et de la flexion, nous avons crée une variable spéciale de caractère générique (RG). Cette variable permet d'établir des règles pour lever l'ambiguïté et pour le pré-traitement concernant la régularisation des amalgames. C'est une variable dont l'ensemble de valeurs est ouvert. Nous avons établi des catégories, sous-catégories pour aider l'analyse morpho-syntaxique, mais ce travail n'est pas exhaustif, nous n'avons pas traité tous les homonymes. Au fur et à mesure que l'étude des anaphores et des conjonctions de subordination, ainsi que la mise à jour de la base LEXIQUE sera approfondie, d'autres variables et valeurs seront certainement définis.

Concernant le modèle de données pour la base de données LEXIQUE, il s'agit d'un modèle relationnel, donc très souple permettant d'inclure des nouvelles variables et des nouvelles tables. Il nous semble que la structure est maintenant stable, la seule chose qui puisse changer, sont des variables et leurs valeurs, c'est-à-dire que nous pourrons créer des nouvelles variables et aussi de nouvelles valeurs de variables déjà existantes. Le module SIDUL doit être un module plus souple aussi, étant donné que la variable RG lui demandera parfois de voir les unités qui ont été prises et aussi de prendre la suivante pour enlever des ambiguïtés éventuelles.

Dans le prochain chapitre nous décrirons la grammaire de reconnaissance et d'extraction des syntagmes nominaux.

« La principale fonction de l'Art est de construire des types sur la base fournie par la Science . » Comte (Auguste), Système de politique positive

Chapitre 10 Grammaire de reconnaissance et

d'extraction des syntagmes nominaux

1 Présentation

Ce chapitre est consacré au développement de la grammaire de reconnaissance et d'extraction des syntagmes nominaux, dans les textes écrits en langue portugaise du Brésil. Elle a été réalisée à l'aide du corpus de syntagmes nominaux extraits dans le cadre de notre mémoire [101] de DEA en Sciences de l'Information et de la Communication 1994/1995. Il s'agit d'un corpus de 6.010 syntagmes nominaux, sans doublons.

Nous présenterons, tout d'abord, la méthodologie adoptée, puis la grammaire proprement dite. Nous présenterons ensuite les conclusions sur le développement de la grammaire de reconnaissance et d'extraction des SN.

2 Méthodologie de développement de la grammaire

Il s'agit essentiellement d'une méthodologie expérimentale, qui a consisté à repérer les SN extraits du corpus dans le cadre de notre cours de DEA et en le décrivant selon la notation d'une grammaire hors contexte. Au tout début de cette tâche, ayant découvert quelques faux SN, nous avons fait une révision de tous les 6010 SN sans doublons, 28 faux SN ont été enlevés. Nous avons donc travaillé avec un corpus de 5.982 SN. Les symboles utilisés pour en faire la description sont ceux des catégories conçues dans le chapitre 9. Nous sommes ainsi partis des vocabulaires terminaux, représentés là par les SN, eux-mêmes, vers la grammaire hors contexte au moyen de la classification des mots. En fait les catégories ont fonctionné comme étant le basculement des vocabulaires terminaux vers les vocabulaires non terminaux.

L'idée initiale a été de trouver les règles les plus stables c'est-à-dire les plus fréquentes, à partir des descriptions de SN. Pour trouver ces règles il a fallu réaliser une procédure de réduction ou de regroupement des composants des descriptions. Cette procédure a été réalisée par le biais de règles de réduction qui regroupent les composants dans les descriptions. La méthodologie peut ainsi être représentée par le schéma de la figure 10.1

3 Etablissement de la Grammaire de Reconnaissance et d'Extraction des SN

Suivant le schéma montré dans la figure 10.1, après le remplacement des unités lexicales

[101] Hélio KURAMOTO. *Maquette d'un système de recherche d'information en utilisant des syntagmes nominaux.* Villeurbanne, 1995. Mémoire du DEA. École Nationale Supérieure des Sciences de l'Information et des Bibliothèques.

par leur catégorie, les 6010 SN ont été réduits à 2587 descriptions de SN. A partir de là, nous avons commencé à chercher des règles qui puissent réduire ce nombre de descriptions. Il est important de ne faut pas oublier que les règles de réduction utilisées par la réduction du nombre de descriptions de SN sont aussi des règles qui appartiennent à la grammaire de reconnaissance et d'extraction des SN.

À partir là, plusieurs révisions ont été faites, soit pour appliquer une nouvelle règle de réduction, soit pour corriger des problèmes d'une application de règle de réduction.

La procédure de réduction de règles a été la plus difficile. Cela à cause de l'ordre d'application des règles de réduction et de la cohérence des résultats. En conséquence, nous sommes contraints de faire, constamment, une évaluation et une révision de tout l'ensemble de descriptions résultant de la procédure de leur réduction. Ce qui est, d'une certaine façon, montré par la figure 10.1.

Toute la procédure d'analyse du corpus de SN et construction du modèle a été faite en utilisant le logiciel Microsoft Access. Celle de réduction de l'ensemble des règles a été facilitée en remplaçant les chaînes de caractères. Ce remplacement a été fait automatiquement et parfois de règle à règle, car il y a eu des situations où un remplacement automatique pouvait entraîner des erreurs. Ainsi il a fallu les remplacer soigneusement.

La notation utilisée est la même adopté pour la grammaire de référence, la BNF, dont les symboles sont :

- D : := <une unité de la catégorie des prédéterminants>
- P : := <une unité de la catégorie des prépositions>
- F : := <une unité de la catégorie des nominaux>
- F_{NOM} : := <une unité de la catégorie nominale, sous-catégorie NA=NOM>
- F_{ADJ} : := <une unité de la catégorie nominale, sous-catégorie NA=ADJ>
- $F_{ADJ,QUA}$: := <une unité de la catégorie nominale, sous-catégorie NA=ADJ, sous-catégorie TA=QUA>
- $F_{ADJ,REL}$: := <une unité de la catégorie nominale, sous-catégorie NA=ADJ, sous-catégorie TA=REL>
- W : := <une unité de la catégorie des adverbes>
- … etc.

Nous décrirons la suite de cette procédure, d'abord en montrant la réécriture de chaque élément d'un SN ensuite la démarche de réduction des règles et l'analyse des problèmes rencontrés dans cette démarche.

3.1 Réécriture des déterminants complexes

Dans le corpus de SN nous avons retrouvé comme déterminant non seulement des unités de la catégorie des prédéterminants, mais aussi des déterminants plus complexes, composés par des prédéterminants et des unités qui appartiennent à d'autres catégories

comme celle de la catégorie F. Les traits de lien de parenté entre la langue portugaise et la langue française apparaissent aussi dans les types de déterminants dans un SN. Ces déterminants ressemblent à ceux traités par Chawk MOHAMAD [102] pour la langue française. Pour distinguer les prédéterminants de ces déterminants composés, nous avons adopté la notation D'. Il y a pourtant quelques différences. Parmi ceux-ci, on peut distinguer les règles concernant les articles partitifs car ce type d'article n'existe pas en portugais.

La syntaxe de ces déterminants complexes (D') se définit comme suit :

- D □ D_{NNU} | D_{NUM}| D_{DEF} | D_{IND}
- D_{NNU} : :=<une unité de la catégorie D, non numérique>
- D_{NUM} : := <une unité de la catégorie D, numérique>
- D_{DEF}: :=<une unité de la catégorie D, défini>
- D_{IND}: :=<une unité de la catégorie D, indéfini>
- D_{NUM}□ dois | três | ... | mil | milhão | ambos | ambas
- E □ E_{INT} | E_{DEC}
- E_{INT} : := <une unité de la catégorie E, chiffre entier>
- E_{DEC} : := <une unité de la catégorie E, chiffre décimal>
- D' □ D_0 | D | E | E + D_{NUM} | D_{DEF} + E + D_{NUM} |
- $D_{Ø}$: := <déterminant zéro ou signale l'absence d'article> [103]

Exemples de déterminants complexes couverts par cette règle :

1. empresas ($D_Ø$)
2. as empresas (les entreprises) (D_{NNU})
3. mil empresas (mille entreprises) (D_{NUM})
4. 500 mil empresas (500 mille entreprises) ($E - D_{NUM}$)
5. as 500 mil empresas (les 500 mille entreprises) ($D_{DEF} - E + D_{NUM}$)

Selon Chawk MOHAMAD il faut établir une contrainte pour cette règle car elle risque d'échouer dans l'analyse de dates, dans la langue française. Pourtant, dans la langue portugaise la date n'est pas toujours précédée d'un article, sauf dans quelques exceptions. On place un article devant une date lorsque cette date relève d'un événement important. Un autre aspect en ce qui concerne la différence entre la date française et la

[102] Chawk MOHAMAD. *La réécriture de D'/les déterminants complexes du français: lexique et syntaxe*. Mémoire de DEA. 1993

[103] l'identification du déterminant zéro sera réalisé par exclusion, en d'autres mots s'il n'y a aucun déterminant selon la règles des déterminants, il faudra voir le mot qui initie la phase pour voir s'il s'agit d'un mot abstrait, ou d'un mot de la classe des noms au pluriel, dans ce cas, on peu dire qu'on a un cas de déterminant zéro.

date portugaise est la syntaxe. En portugais on exprime la date sous la forme : 99 de mois de 9999.

Exemple : 20 de janeiro de 1996.

Ainsi, le repérage du syntagme nominal sous la forme de date doit être fait selon une procédure spécifique et non pas selon la règle normale du syntagme nominal. Cette procédure doit suivre une règle de la forme :

<date>
□EINT + P-DE + <mois> + P-DE + EINToù<mois> □ Janeiro | Fevereiro |Março | Abril | Maio | Junho | Julho | Agosto | Setembro | Outubro | Novembro | Dezembro

La procédure veille à la valeur du premier E_{INT} car les jours peuvent être dans la fourchette de 1 a 30 pour les mois d'Avril, Juin, Septembre, Novembre ; et dans la fourchette de 1 a 31 pour les mois de Janvier, Mars, Mai, Juillet, Août, Octobre et Décembre ; et de 1 a 28 ou 29 selon l'année pour le mois de février, 29 pour les années bissextiles et 28 pour les autres.

D'autres formes de déterminant complexe ont été trouvées dans le corpus de SN.

Exemple :

a maior parte da força de trabalho (la plus grande partie de la force de travail) 1.

a maioria das novas atividades (la majorité des nouvelles activités) 2.

Ces déterminants ressemblent à ceux traités par Chawk MOHAMAD. On peut établir une règle de réécriture pour ces déterminants en examinant les exemples suivants :

a maior parte da força de trabalho (la plus grande partie de la force de travail) D' □DNNU + A + N + P-DE + DNNUoù A□FADJ,QUA N □FNOM P-DE : := <la préposition DE> 1.

a maioria das novas atividades (la majorité des nouvelles activités) D' □ DNNU + N + P-DE + DNNU 2.

a baixa capacitação técnico-científica (la faible compétence technico- scientifique) D' □ DNNU + A 3.

D'après ces exemples et sachant qu'un N peut être réécrit comme A + N, ce qui ne change pas le statut de N, il demeure un N, c'est-à-dire un prédicat libre. Nous pouvons établir les règles suivantes :

D'
□ DNNU + N + P-DE + DNNU | DNNU + N + P-DE | DNNU + A

Il existe des mots dans la classe F qui sont des homonymes car ils peuvent parfois jouer un rôle de déterminant et parfois de non déterminant dans une autre construction syntaxique. En fait la question qu'on se pose est celle-ci : faut-il vraiment prend en compte ce genre de déterminant ? Que ces expressions (exemples 1, 2 et 3) jouent le rôle de

déterminant, il n'y a aucun doute. La solution, donnée par Chawk MOHAMAD — de créer une variable, accompagnée de règle de contrainte, permettant de distinguer d'une part les noms pouvant entrer dans la formation de déterminants complexes, d'autre part ceux qui ne le peuvent pas — nous semble régler ce problème. Or, la question est en outre de savoir si cela vaut la peine de mettre cette information pour chaque unité lexicale, dans la base de données LEXIQUE. Cela implique une augmentation de tâche dans la procédure de saisie d'information pour chaque unité lexicale. Et quelques lignes supplémentaires de programmes dans le module d'analyse seront nécessaires pour vérifier les règles de contraintes.

En fait, il s'agit d'un alourdissement de la procédure d'analyse morpho-syntaxique. Du point de vue de la reconnaissance de SN, il y a déjà le prédéterminant, ce qui donne déjà l'indication de début du SN. Ainsi le fait de ne pas prendre en compte des déterminants complexes ne doit pas déranger la procédure de reconnaissance de SN et de leur extraction. Du point de vue de l'indexation, il nous semble aussi qu'il n'y a pas de problèmes, puisque les SN seront indexés quand même. Le seul inconvénient est qu'on aura un niveau de plus dans la structure des syntagmes nominaux.

Prenons l'exemple (1) :

A maior parte da força de trabalho (La plus grande partie de la force de travail). SN_1 : *a força de trabalho* (la force de travail) SN_2 : *a maior parte da força de trabalho* (la majeure partie de la force de travail)

Si la procédure de reconnaissance et d'extraction de SN prenait en compte les déterminants complexes ce SN serait de premier niveau au lieu de deuxième niveau selon l'exemple ci-dessus. La conséquence de l'adoption de ne pas prendre en compte les déterminants complexe est l'utilisation d'un peu plus d'espace en disque dur. Or, la saisie d'information pour chaque unité lexicale pour rendre faisable la reconnaissance des déterminants complexes peut prendre beaucoup plus d'espace en disque puisqu'il faut mettre ces informations pour chaque unité lexicale.

Nous pouvons donc envisager trois solutions possibles pour les trois types de déterminants complexes (ceux des exemples 1,2 et 3), soit nous ne prenons pas en compte l'existence de ces déterminants, mais seulement du prédéterminant, soit nous prenons en compte et créons les variables nécessaires, ou bien nous mettons tous les déterminants complexes dans la base de données LEXIQUE.

La première solution proposée peut présenter l'avantage que nous avons exposé plus haut. Etant donné que le but de ce travail est l'indexation automatique, il nous semble que cette solution ne compromet pas l'indexation elle-même et ni la recherche d'information.

La deuxième solution serait idéale. Elle exige des efforts dans la mise à jour de la base de données et dans la programmation de l'analyseur. Pour l'indexation nous ne cherchons pas l'identification des déterminants complexes mais les syntagmes nominaux. C'est une bonne solution pour des travaux comme la traduction automatique ou d'autres applications qui exigent l'identification de ces éléments.

En ce qui concerne la troisième solution, elle peut résoudre de manière satisfaisante

le problème, mais elle est quand même ennuyant étant donné qu'on va surcharger la base de données LEXIQUE.

Ainsi, la décision de ne pas repérer les déterminants complexes nous semble bonne solution en ce qu'elle prend l'indexation comme but la reconnaissance et l'extraction des SN.

Encore :

mais ou menos 3000 anos [environ 3000 ans] D' □ WQUA + DNUM WQUA ::= <élément de la catégorie W, sous-catégorie QUA>	1.
mais de 3000 anos [plus de 3000 ans] D' □ WQUA + P-DE + DNUM	2.
cada 10-15 anos D' □ DIND + I I ::= <fourchette numérique> I □ EINT -EINT | EINT	3.
Toda a empresa [Toute l'entreprise] D' □todo o | toda a | todos os | todas as (tout le | toute la | tous les | toutes les) En effet, cette règle de réécriture doit être exprimée ainsi : D' □ DIND + DDEF Cependant, il faut créer dans la variable RG pour les unités TODO, TODA, TODOS, TODAS, une valeur (TOD) avec la force d'une règle de contraint, en disant que ces unités peuvent être suivies d'un déterminant défini du type "o, a, os, as"	4.

Le mot *todo* (*tout*) suivi d'un article, au singulier, détermine la totalité du substantif. Exemple : Toda a empresa [toute l'entreprise] (c'est-à-dire l'ensemble complet d'une entreprise donnée).

Tandis que le mot *todos* suivi d'un article au pluriel, fait référence à l'ensemble de tous les objets représentés par le substantif qui suit. Exemple : Todos os franceses [Tous les français] (c'est-à-dire l'ensemble de tous les français et non seulement un français).

Par contre, il faut signaler qu'il y a des mots qui refusent l'utilisation de l'article. Exemple : Todo Portugal pensa assim [Tout le Portugal pense comme cela).

L'utilisation du mot *todo* sans être suivi d'un article donne un sens de généralité au substantif qu'il précède. Exemple : Toda casa deve ser reformada [Toute maison doit être réparée] (c'est-à-dire une maison quelle qu'elle soit doit un jour être réparée).

Nous allons ainsi faire une synthèse de la règle des déterminants.

D'

□ DØ | D | E | E + DNUM | DDEF + E + DNUM| WQUA + DNUM | WQUA + P-DE + DNUM | DIND + I | DIND + DDEF

D

□ DNUM | DNNU | DDEF | DIND

3.2 Réécriture des nominaux et d'autres constituants d'un syntagme nominal

Après avoir établi la description des déterminants, nous avons établi la description des

nominaux. Tout d'abord nous avons défini :

N0
::= <centre du syntagme nominal>

N0
□FNOM | FNAN

N
: := <prédicat libre>

N
□ N0

Nous avons défini le symbole N_0 pour désigner le centre de syntagme nominal, qui sera toujours une unité de la catégorie F_{NOM} et le symbole N pour indiquer qu'il s'agit d'un prédicat libre. Mais, dans ce cas, N peut être composé non seulement d'un N_0, mais par d'autres éléments comme les F_{ADJ} et demeurer un prédicat libre. Nous reviendrons sur ce point.

Remarque : Dans la procédure de description des SN, nous avons remplacé les F NOM directement par N pour une question de simplification de la tâche de remplacements des vocabulaires terminaux par ceux des non-terminaux, étant donné que nous savions déjà qu'un N peut être dans sa forme la plus simple un N_0.

Par définition, nous avons pris directement les noms propres et les sigles comme des syntagmes nominaux. Ils sont représentés par le symbole N''.

N''
: := <syntagme nominal>

Selon la démarche du groupe SYDO, et étant donné que nous avons défini l'absence de l'article par le symbole $D_{Ø}$, nous pouvons maintenant représenter le syntagme nominal comme :

N''
□D' + N' où :N' : := <prédicat lié> Et la forme la plus simple d'un prédicat lié est la composition suivante :N' □ N

Cependant le prédicat lié peut avoir d'autres compositions plus complexes. Nous allons le montrer plus loin.

Les adjectifs sont représentés par le symbole A.

A
: := <adjectifs de qualité>

A
□ FADJ, QUA | FNAN,QUA

A0
: := <adjectifs de relation>

A0
□ FADJ,REL

Les prépositions ont été représentées par le symbole P. Envisageant de connaître le rôle des prépositions dans le syntagme nominal, nous avons adopté la démarche de les réécrire dans la forme suivante :

P
: := <prépositions simples>

Nous avons d'abord repéré les prépositions en utilisant le symbole de leur catégorie en ajoutant la préposition elle-même. Cette procédure a été utilisée dans l'intention de connaître les prépositions et les locutions prépositionnelles trouvées dans le corpus.

P-XX où XX était la préposition que ce symbole remplace dans la réécriture des SN. Dans une deuxième analyse du corpus, nous avons remplacé toutes les prépositions de la forme P-XX, par P'.

P'
: := <préposition complexe>

P'
P | P0 oùP : := <prépositions simples>P0 : := <locution prépositionnelle>Remarquons que dans ce modèle, les deux types d'unités sont mis dans la même catégorie lexicale. Nous mettons deux symboles ici seulement afin de distinguer les deux types d'unités. C'est une ouverture pour l'avenir, envisageant la possibilité de formalisation d'une grammaire pour décrire les locutions prépositionnelles et de ne pas les mettre dans le lexique. Ainsi, la séparation en deux symboles et la création d'un symbole P', préposition complexe pour représenter les deux types de prépositions est, dans ce modèle, purement symbolique.

P
 DE | EM | COM | PARA | SOB | SOBRE | POR | A | SEM | …

P0
 graças a | a respeito de | acima de | abaixo de | ...

Nous avons remplacé les unités de la catégorie des adverbes par leur symbole W.

W
::= <les adverbes>

W
 mais | bastante | apenas | pouco | nunca | não | … (plus | assez | seulement | peu | jamais | ne pas)

Les locutions adverbiales ont été remplacées par W'.

W'
::= <les locutions adverbiales>

W'
 quanto a | até mesmo | muito pouco |…

W'
 W | <locutions adverbiales>

Les verbes, fléchis ou à l'infinitif ont été remplacé par la lettre V.

V
::= <unité de la catégorie des verbes>

Les conjonctions ont été remplacées par la lettre C, suivie d'un tiret et la conjonction elle-même.

C
: := <conjonctions>

C-XX où XX est la conjonction elle-même. Cette procédure a été adoptée car nous étions encore dans la procédure de repérage des éléments et les catégories n'étaient pas encore totalement définies. Après quelques définitions, nous avons remplacé la notation C-XX par C lorsque XX était égale à *E* (et), *OU* (ou). Ainsi que nous avons remplacé les conjonctions utilisées dans l'énumération de mots ou des unités plus complexes et éventuellement des phrases, par le symbole C, de la catégorie des conjonctions de coordination. Nous avons laissé les autres conjonctions sous la forme C-XX.

Les pronoms relatifs ont été remplacés par le caractère R, la catégorie des conjonctions de subordination n'ayant pas été encore définie. Pourtant, dans la présentation d'un extrait de la dernière description de SN qui sera présenté plus loin dans ce chapitre, nous avons remplacé les pronoms relatifs R, du type "que", par Q_{QUE} . Nous avons adopté la même démarche pour les conjonctions de subordination (Q) indiquant dans la forme d'indice la conjonction elle-même. Exemple : Q_{comme}.

3.3 Procédure d'établissement de règles

Ayant défini les symboles pour la réécriture de chaque élément d'un SN, nous avons fait le remplacement de chaque symbole du vocabulaire terminal pour son symbole correspondant dans le vocabulaire non terminal. Ces symboles étaient plutôt ceux des catégories des unités lexicales. Auparavant ces symboles correspondaient à ceux des catégories (F_{NOM}, $F_{ADJ,QUA}$, $F_{ADJ,REL}$, P, W, etc).

Ainsi, après cette première phase, nous avons eu 2587 descriptions de syntagmes nominaux différents.

La deuxième phase consistait à régulariser ces représentations, en tenant compte des déterminants complexes (D'), les déterminants zéro inclus, en remplaçant le F_{NOM} par N, $F_{ADJ,QUA}$ par A, $F_{ADJ,REL}$ par A_0, P-XX par P', C-E C-OU par C. Dorénavant nous avons adopté le terme déterminant complexe pour désigner une classe plus grand qui enveloppe tous les types de déterminants définis dans ce modèle. À partir de ces remplacements préliminaires nous avons commencé à établir effectivement la grammaire de reconnaissance et d'extraction des syntagmes nominaux. En fait ces remplacements préliminaires correspondent au passage du vocabulaire terminal au vocabulaire non terminal.

Comme pour la langue française, dans la langue portugaise le syntagme nominal a comme règle générale de réécriture :

N''

□ D' + N' (1) oùN'': := <syntagme nominal>D' : := <déterminant complexe>N' : := <prédicat lié>

Nous avons observé qu'au fur et à mesure que nous faisions la réduction des descriptions des SN en appliquant les règles de réduction, le nombre de descriptions diminuait et la fréquence d'occurrence de la description générale d'un SN, $(D' + N')_{N''}$, augmentait. Ce phénomène nous amène à la conclusion que l'application des règles de réduction fait converger les descriptions vers la règle générale de réécriture des SN (N'' □ D' + N') et que les règles de réduction font partie de la grammaire de reconnaissance et d'extraction des SN. Celle-ci est donc constituée de la règle générale de réécriture des SN et des règles de réduction trouvées dans la procédure d'analyse et d'établissement de ses règles.

Ces règles ne peuvent être appliquées d'une manière quelconque, mais selon une procédure systématique, soigneuse et dans un certain ordre.

C'est cette procédure que nous allons montrer en regroupant les règles selon l'ordre logique d'application.

3.3.1 Constitution du vocabulaire non terminal A (les prédicats adjectivaux)

Tout d'abord, nous essayons de réduire les adjectifs composés par les adjectifs eux-mêmes et les adverbes, avant la réduction des suites coordonnées d'adjectifs.

1. A □ A + WAAJ | A + W'AAJ | WAAJ + A | W'AAJ + A Nous avons trouvé plusieurs SN avec la combinaison A + W ou W + A. Exemple : ambientes mais remotos (des environnements les plus remotes) On le réécrit : DØ + N + W + A Etant donné que : D' □DØ Et que : A □W + A On réduit cette réécriture à : D' + N + A D'autres exemples de A + WAAJ : necessidades bem idenficadas (des besoins bien identifiés) ; características muito próprias (des caractéristiques très propres) ; os três fatores tradicionalmente relacionados (les trois facteurs traditionnellement liés). Nous voyons dans ces exemples que les adverbes sont utilisés pour renforcer une qualification donnée, c'est-à-dire l'adjectif de qualité. Ce sont normalement des adverbes qui modifient les adjectifs (AAJ). Ce qui nous a amené à les adopter comme partie du vocabulaire non terminal A.
2. A □ A + <,> + A | A + C + A où [C □ e |ou] Une suite coordonnée d'adjectif de qualité (A, A et A) peut être représentée par un seul A. Exemple : a informação tecnica, científica e econômica (L'information technique, scientifique et économique) Cela se réécrit : D' + N + A + <,> + A + C + A Appliquant la règle A □ A + <,> + A, cette description est réécrite comme : D' + N + A + C + A Appliquant A □ A + C + A, cette description est réécrite comme : D' + N + A
3. A' □ A' + <,> + A' | A' + C + A' où : Cette règle de réécriture concerne à la coordination d'adjectifs de relation. Exemple : as atividades agrícolas, industriais ou artesanais (les activités agricoles, industrielles ou artisanales) Cela se réécrit comme : D' + N + A' + <,> + A' + C + A' Appliquant A' □ A' + <,> + A', la description peut être

réécrite comme : D’ + N + A’ + C + A’ Appliquant A’ □ A’ + C + A’, la description ci-dessus peut être réécrite comme : D’ + N + A’

3.3.2 Constitution du vocabulaire non terminal N (les prédicats nominaux)

Après la réduction des adjectifs nous avons travaillé dans la réduction des N. Avant de discuté ce genre de réduction, il faut rappeler la composition d’un N : N □N_0

1. N □ N + <,> + N | N + C + N où C □e | ou Exemple : a análise, interpretação, avaliação e comunicação da informação pelos meios convenientes (L’analyse, l’interprétation, l’évaluation et la communication de l’information par les moyens appropriés) La description de ce SN est : D’ + N + <,> + N + <,> + N + C + N + P + D’ + N + P + D’ + N + A Appliquant N □ N + <,> + N, la description peut être réécrite comme : D’ + N + <,> + N + C + N + P + D’ + N + P + D’ + N + A Appliquant encore la même règle, nous obtiendrons la description suivante : D’ + N + C + N + P + D’ + N + P + D’ + N + A Enfin, appliquant la règle N □ N + C + N, nous arrivons à la description suivante : D’ + N + P + D’ + N + P + D’ + N + A Après quelques essais sur le corpus de SN, nous avons appris que le bon ordre était d’abord la réduction des A, puis des N et c’est seulement après cela que nous pouvons faire d’autres réductions comme l’utilisation des règles des syntagmes prépositionnels et d’expansions prépositionnelles. Autres exemples de suite coordonnés de N :

- *a armazenagem, recuperação e utilização de documentos/informações nas organizações* (le stockage, la récuperation et l’utilisation de documents/informations dans les organisations) ; Là, il faut d’abord faire un pré-traitement du signe « / », puisque ici il a la valeur « ou », donc il faut le remplacer par la conjonction « ou ». Et seulement après, faire la procédure de réduction en appliquant les règles de réduction de suites de coordonnés de N.
- *a classificação, organização e recuperação de informações* (la classification, l’organisation et la récuperation d’informations)
- *a implantação, manutenção e aperfeicoamento das unidades de informação* (L’implantation, l’entretien et le perfectionnement des centres d’information).

En ce qui concerne les suites coordonnées, soit d’adjectifs, soit de noms, il nous semble important non seulement de prendre le SN complet, avec les suites coordonnées, mais de faire un pré-traitement de manière à prendre les syntagmes nominaux formés à partir de chaque nom ou adjectif de la suite coordonnée. Dans la mesure où la proposition de cette recherche est de construire un Système de Recherche d’Information Assisté par Ordinateur, où nous allons privilégier l’interaction homme-machine, il faut présenter tous les syntagmes nominaux possibles ; même qu’on propose parfois de faux syntagmes nominaux, l’utilisateur saura distinguer les bons SN. Il nous semble que cette solution est préférable car elle évite l’alourdissement du système à cause des algorithmes de levée d’ambiguïté. Sachant, encore, au préalable, qu’assez souvent nous aurons du mal à trouver la bonne solution ou les bons syntagmes nominaux.

Il est vrai que les travaux menés par Omar LAROUK dans sa recherche des suites coordonnées pourraient aider à résoudre les problèmes de coordination de noms, adjectifs et même de syntagmes nominaux. Mais, nous avons décidé d'adopter une solution plus simple en prenant en compte les caractéristiques de notre proposition de système de recherche d'information.

Sur la coordination de deux N, comme nous l'avons déjà signalé, il a fallu veiller à la procédure de remplacement d'une telle suite. L'exemple suivant peut montrer un des problèmes rencontrés dans cette procédure :

A relação entre biblioteconomia e ciência da informação (Le rapport entre la bibliothéconomie et la science de l'information.) On réécrit ce syntagme nominal sous la forme suivante : D' + N + P + N + C + N + P + D' + N

Nous sommes tentés de faire, d'emblée, deux sortes de réduction D' + N pour N'' étant donné que N est un cas particulier de N' et dans la mesure où D' + N' décrit un syntagme nominal, nous pourrions donc le remplacer par N''. L'autre réduction était N + C + N, car la coordination de deux nominaux a comme résultat un nominal (N). Pourtant, cette deuxième règle ne peut pas être appliquée d'emblée puisque nous ne pourrions pas réduire simplement Bibliothéconomie et science. Le mot science a un complément nominal (information) qu'il faut prendre en compte. Il faut donc réduire l'expression *ciência da informação* (science de l'information) avant la résolution de la coordination des nominaux N + C + N. Ainsi, nous arrivons à :

- D' + N + P + N + C + N + P + (D' + N')=N'' ou
- D' + N + P + N + C + N + P + N''
- Mais, P + N'' est un syntagme prépositionnel (SP), donc :
- D' + N + P + N + C + N + SP
- Et la construction N + SP est un prédicat lié (N') :
- D' + N + P + N + C + N'

Nous rencontrons là un autre problème, en fait ce qui était une coordination entre deux N, était en réalité la coordination entre un N et un N'. Une fois encore, cette constatation nous a forcé à augmenter le soin dans la procédure de réduction des règles. Mais ici, la question qu'on se pose est de savoir si l'on peut coordonner deux éléments de niveaux différents.

Omar LAROUK [104] reprend une définition de G. ANTOINE, qui à son tour a repris la définition de BOËSE qui dit que ***« La coordination renferme une série de termes situés sur le même plan »*** .

Une autre définition par la coordination a été donnée par G. ANTOINE : ***« On a une coordination lorsqu'il y a mise en ordre de deux termes (membres) ou davantage équilibrés ou harmonisés dans un ensemble créant entre eux une unité relative »*** [105] .

[104] Omar LAROUK. *Extraction de connaissances à partir de documents textuels : traitement automatique de la coordination (connecteurs et ponctuation).* Thèse de doctorat. Lyon : Université Claude Bernard – Lyon I, 1993. p. 56.

Maurice GREVISSE (1986, p. 382) définit : ***« La coordination est la relation, explicite ou implicite, qui unit des éléments de même statut : — soit des phrases, soit, à l'intérieur d'une phrase, des termes qui ont la même fonction par rapport au même mot »*** [106] .

Ces interprétations nous ont amenés à la réflexion suivante : est-ce que le terme « science de l'information » est-il une expression figée comme la *rose des vents* ?

M. LE GUERN rappelle que dans le terme *la rose des vents*, malgré l'existence d'un article après la préposition, on ne retrouve plus le syntagme « les vents ». Selon lui, ***« la lexicalisation de l'expression fait que l'article les (des = de + les) ne porte plus ici le présupposé d'existence. »*** [107] . Selon lui, le résultat de cette lexicalisation fait que l'objet désigné par le syntagme figée n'a rien à voir avec les mots qui le composent. Dans ce cas, le syntagme nominal désigne un objet qui n'a rien à voir ni avec la rose ni avec le vent. En ce sens, nous ne pouvons pas faire la même analogie avec le terme « science de l'information », car dans ce terme existe le rapport avec la science et avec l'information.

Émile BENVENISTE appelle ce genre d'expression comme des synapses. Lesquelles selon lui, elles sont caractérisées par : ***« 1° la nature syntaxique (non morphologique) de la liaison entre les membres ; 2° l'emploi de joncteurs à cet effet, notamment 'de' et 'à' ; 3° l'ordre déterminé + déterminant des membres ; 4° leur forme lexicale pleine, et le choix libre de tout substantif ou adjectif ; 5° l'absence d'article devant le déterminant ; 6° la possibilité d'expansion pour l'un ou l'autre membre ; 7° le caractère unique et constant du signifié. »*** [108] .

Nous voyons que le terme « science de l'information » ne peut pas être considéré comme une expression figée ou une synapse car il y a un article devant le mot déterminant (information). Il ne présente pas la septième caractéristique citée plus haut, comme l'avait déjà signalé M. LE GUERN.

D'autre part, nous pouvons supposer que le terme « bibliothéconomie » ait un statut de N' car il s'agit d'un nom abstrait ; de plus il est aussi le nom d'un domaine de connaissance. Et, selon ce que nous avons vu dans le chapitre 7, les noms propres et les noms abstraits ont le même comportement, ils désignent un objet précis dans le discours, donc ils n'exigent pas l'article. Il nous semble donc raisonnable de prendre le terme bibliothéconomie comme un syntagme nominal avec déterminant zéro. Et, en tenant en compte les règles d'utilisation des articles dans la langue portugaise, en cas de coordination des éléments de même nature, nous pouvons aussi considérer le terme « science de l'information » aussi comme un syntagme nominal avec déterminant zéro. Ainsi, il ne s'agit pas ici d'une coordination de deux prédicats libres, mais de deux

[105] Citation extrait de la thèse d'Omar LAROUK, p. 57

[106] Maurice GREVISSE. Le Bon Usage. 12ème édition. Editions DUCULOT, 1986. p. 382.

[107] Michel LE GUERN. « Un analyseur morpho-syntaxique pour l'indexation automatique ». IN. : *Le français Moderne.* p. 31.

[108] Émile BENVENISTE. *Problèmes de linguistique générale, 2.* Collection Tel. Paris : Editions Gallimard, 1974, p. 172-173.

syntagmes nominaux, ce qui donne la description suivante :

- D' + N + P + (D' + N') + C + (D' + N')
- Donc,
- D' + N + P + N'' + C + N''

Cet exemple nous a montré le soin à apporter à l'application des règles de réduction qui ne peut se réduire de manière automatique à une simple opération couper-coller. De plus, cette analyse renforce l'importance de caractériser les noms abstraits comme tels. C'est-à-dire que nous justifions par-là l'importance de la variable VN laquelle indique si un F_{NOM} est un nom concret ou abstrait.

Ayant montré le problème principal de la méthodologie adoptée dans cette recherche, il faut ajouter que les règles ne sont pas appliquées une seule fois sur le corpus de SN, mais plusieurs fois, car après l'application successive de ces règles, de semblables descriptions peuvent réapparaître.

À ce moment là, certaines descriptions ou même partie de descriptions pouvaient être regroupées comme des syntagmes nominaux puisqu'il y avait des descriptions sous la forme :

D' + N et D' + N'

Etant donné que N est un cas particulier de N' nous les avons, dans ce cas, remplacés par N' et ensuite nous avons appliqué, sur tout le corpus de SN, la règle générale :

N''
 D' + N'

3.3.3 Constitution des Syntagmes prépositionnels et Expansions prépositionnelles

La démarche consiste à regrouper les P, soit les prépositions simples, soit les prépositions constituées de locutions prépositionnelles sur la désignation P'.

Ensuite, nous avons regroupé les syntagmes prépositionnels et les expansions prépositionnelles, de la façon suivante :

SP
 P' + N''

EP
 P' + N

Les règles de réécriture des syntagmes prépositionnels et des expansions prépositionnelles établis pour la langue française sont aussi valables pour ceux respectifs de la langue portugaise. Or, nous avons remarqué une quantité important de SN sans déterminant ou ce qu'on appelle avec déterminant zéro. Ce phénomène provoque une ambiguïté entre l'expansion prépositionnelle et le syntagme prépositionnel lorsque le nom

qui suivit la préposition est un nom abstrait. Exemple :

- os sistemas de informação (les systèmes d'information)
- Ce SN est réécrit comme :
- D' + N + P ' + N.

Normalement nous prenons la séquence P + N comme une expansion prépositionnelle (EP). Cependant, le N qui suit la préposition « de » est un nom abstrait. Dans ce modèle que nous venons de concevoir, un nom abstrait est une marque d'existence d'un déterminant zéro. C'est-à-dire que dans ce modèle, le SN « os sistemas de informação » sera pris comme un SN de deuxième niveau, puisqu'il est composé d'un syntagme prépositionnel au lieu d'une expansion prépositionnelle. Ainsi, en suivant ce modèle, ce SN serait réécrit comme :

- D' + N + P' + DØ + N
- En appliquant la règle D' □ DØ, on obtient la description suivante :
- D' + N + P' + D' + N
- Sachant que N est un cas particulier de N', nous pouvons remplacer le dernier N :
- D' + N + P' + D' + N'
- En appliquant la règle N" □D' + N", on obtient la description suivante:
- D' + N + P + N''
- En appliquant la règle SP □ P' + N"
- D' + N + SP

Nous avions un syntagme prépositionnel au lieu d'une expansion prépositionnelle. Or, dans le SN « os sistemas de informação », « de informação » ne joue pas le rôle d'un syntagme prépositionnel. Ce SN a une configuration pareille à celle du SN « Le placard de cuisine ». Là, aussi bien la suite « de cuisine » que « de informação » jouent le rôle d'un qualificateur et non le rôle de déterminant. Les composants "de cuisine" selon M. LE GUERN, dans son article de la revue Le Français Moderne, jouent le rôle d'une EP et non d'un SP.

M. LE GUERN rappelle que, dans ce cas, si le mot « information » ne peut pas être référencié par un élément anaphorique après son apparition dans le discours, il s'agit, en fait, d'une expansion prépositionnelle. Et, cela est le cas, nous ne pouvons pas faire référence au mot « information » à travers un élément anaphorique.

Autres exemples de ce même genre de problème :

1. os serviços de informação (les services d'information)
2. a atividade de marketing (l'activité de marketing)
3. a atividade de planejamento estratégico (l'activité de planification stratégique)
4. o processo de inovação tecnológica (la procédure d'innovation technologique)

a capacidade de percepção (la capacité de perception) 5.

a demanda de informação (la demande d'information) 6.

o direito de propriedade (le droit de propriété) 7.

a base de dados Energyline 8.

15000 empresas de manufatura (15000 entreprises de manufacturation) 9.

a abordagem de custos de serviços de informação (l'approche des coûts des services d'information) 10.

os fatores de produção clássicos (les facteurs classiques de production) 11.

os recursos de informação (les ressources d'information) 12.

a agregação de valor (l'ajout de valeur) 13.

a análise de conteúdo (l'analyse de contenu) 14.

Il y a trois solutions possibles pour régler ce problème : 1) en prenant en compte le fait que le mot abstrait présuppose un déterminant zéro, considérer les suites P + N comme, en fait, des suites P + $D_{Ø}$ + N, et ainsi, la dégrée du SN qui porte ce type de suite augmentera d'un niveau, car ce qu'on prenait comme des expansions prépositionnelles sera pris comme des syntagmes prépositionnels ; 2) en analysant les exemples présentés, nous pourrions adopter la mesure de ne prendre comme des SN sans déterminant que les noms abstraits au pluriel ; 3) mettre dans le lexique tous ces termes, car ils sont des mots composés.

La solution (1) n'est pas acceptable car elle cache de vraies expansions prépositionnelles derrière de faux syntagmes prépositionnels. Parmi les exemples montrés plus haut, presque tous les SN ne contiennent que des expansions prépositionnelles, sauf les exemples 8 et 10. Dans l'exemple 8 le SN « as bases de dados » un nom concret au pluriel suit la préposition, donc nous avons là un syntagme prépositionnel. Dans l'exemple 10 « *a abordagem de custos de serviços de informação* » (l'approche de coûts de services d'information)], il nous semble que les SN « custos de serviços de informação » et « serviços de informação » sont des vrais SN avec déterminant zéro. Cependant la suite « de informação » du SN « serviços de informação » n'est pas un syntagme prépositionnel mais une expansion prépositionnelle.

La deuxième solution nous semble plus approprié car elle garantit au moins la prise en compte de la plus grande partie des expansions prépositionnelles. Le problème qu'elle pose est que quelques expansions prépositionnelles risquent d'y échapper à cause du fait que le mot qui suit la préposition est un nom abstrait au pluriel.

La troisième solution ne nous paraît pas devoir être retenue car la taille du lexique peut s'agrandir excessivement.

Il existe un autre problème lié à la question de l'ordre d'application des règles de EP et de SP. Prenons comme exemple une suite de EP :

- a fase de pesquisa de opiniao (la phase de recherche d'opinion)
- Dont la réécriture est :

- D’ + N + P + N + P + N,

Il y a deux façons de traiter cette réécriture : soit nous faisons le remplacement de la suite P + N par EP dans la direction de gauche à droite, soit nous le faisons à partir de la dernière suite P + N. Le résultat sera toujours le même.

- D’ + N + P + N + P + N => D’ + N + EP + EP
- et maintenant, nous pouvons remplacer la suite N + EP par N :
- D’ + N + EP => D’ + N
- Si nous faisons la réduction de la description à partir de la dernière suite P + N, nous aurons :
- D’ + N + P + N + EP
- et nous savons que N + EP est aussi un N, donc :
- D’ + N + P + N
- étant donné que N + EP est un N, on peut la réécrit comme :
- D’ + N + EP
- et finalement, encore en appliquant la même règle, la description finale est réécrit :
- D’ + N

La différence entre ces deux approches de traitement de EP est due au fait que la deuxième approche tient compte de la valeur sémantique des termes qui sont dans le SN. Tandis que la première approche ne fait aucune attention à la signification des termes qui sont dans le SN, c’est seulement l’ordre adopté pour l’analyse de textes qui compte. C’est comme dans l’évaluation d’une ligne de command d’un programme d’ordinateur. En ce cas, la deuxième approche a tenu compte du fait que *la phase* concerne la *recherche d’opinion* et non seulement la *recherche*. Ainsi, la dernière approche a résolu d’abord la réduction de la réécriture du terme « recherche d’opinion » qui est en fait un N, étant donné qu’il est un mot composé. Et seulement après, nous avons traité la première partie « la phase de » avec le résultat de la première réduction (N – recherche d’opinion). Ainsi, il nous semble que la dernière approche serait plus correcte du point de vue de la cohérence entre la réécriture et les termes qui se trouvent dans les SN.

Nous pouvons montrer d’autres exemples de ce genre de réécriture :

1. método de controle de qualidade - (méthode de contrôle de qualité) DØ + N + P' + N + P' + N.
2. o conceito de economia de rede - (le concept d'économie de réseau) D' + N + P' + N + P' + N.

Cependant, en adoptant cette dernière approche comme le bon ordre de remplacement de la suite P + N par EP, cela pose de problème lorsqu’il y a combinaison d’une suite de EP et SP. Ce que nous montrons à travers l’exemple suivant :

- o conceito de administração dos recursos de informação

- (Le concept d'administration des ressources d'information)
- Dont la réécriture est la suivante :
- D' + N + P' + N + P' + D' + N + P' + N
- En appliquant EP □P' + N, la description est réécrite comme suit :
- D' + N + P' + N + P' + D' + N + EP
- En appliquant N □N + EP, la description est ainsi réécrite :
- D' + N + P' + N + P' + D' + N
- En appliquant N' □ N, la description est réécrit comme suit :
- D' + N + P' + N + P' + D' + N'
- En appliquant N'' □D' + N', la description est réécrite comme suit :
- D' + N + P' + N + P' + N''
- En appliquant SP □ P' + N'', la description est réécrite comme suit :
- D' + N + P' + N + SP
- En appliquant N' □ N + SP, la description est réécrite comme suit :
- D' + N + P' + N'

Le résultat n'est pas acceptable car la suite P' + N' ne peut pas exister. Cependant, si nous adoptons la démarche de traiter la réécriture de gauche à droite sans prendre en compte la signification des termes qui sont dans le SN, nous arrivons à un résultat correct.

- D' + N + P' + N + P' + D' + N + P' + N
- En appliquant EP □P' + N, la description est réécrite comme suit :
- D' + N + EP + P' + D' + N + P' + N

Le traitement de gauche à droite est fait s'il n'y a pas de déterminant. Si l'on trouve un déterminant, il faut d'abord résoudre les éléments qui suivent le déterminant puisqu'il s'agit d'un nouveau syntagme nominal. Dans ce nouveau SN, le traitement obéit aussi au sens de gauche à droite. C'est-à-dire que dans la mesure où les déterminants sont des marques de début de SN, ils changent l'ordre de priorité de traitement des éléments dans un SN. Ainsi, la dernière description signale l'existence d'une EP à la fin du SN, la description est donc réécrite à cause de l'application de la règle de l'EP, comme suit :

- D' + N + EP + P' + D' + N + EP
- En appliquant N □ N + EP, la description peut être réécrite ainsi :
- D' + N + EP + P' + D' + N
- En appliquant N' □ N, la nouvelle réécriture de la description est :
- D' + N + EP + P' + D' + N'
- Sachant que D' + N' décrit un SN, nous pouvons réécrire la dernière description comme suit :
- D' + N + EP + P' + N''

- En appliquant SP □ P' + N'', la description est réécrite comme suit :
- D' + N + EP + SP
- Sachant que N □ N + EP, la description est réécrite comme suit :
- D' + N + SP
- Et, finalement se N' □ N + SP, la description devient :
- D' + N'

Cet exemple montre, qu'outre le fait de traiter soigneusement la réduction des descriptions des SN, l'ordre naturel de traitement est pris de gauche à droite, et que le fait de trouver un déterminant modifie l'ordre de traitement, d'abord parce qu'il s'agit du début d'un nouveau SN, il faut donc traiter les éléments qui le composent. Pourtant, dans ce nouveau SN, l'ordre de traitement suit l'ordre général, c'est-à-dire de gauche à droite dans la mesure où il n'y a pas de déterminant.

3.3.4 Réduction des descriptions du type N + A et A+ N

Les vocabulaires N, EP et SP ayant été consolidés, nous étions prêts pour traiter les suites N + A et A + N. Normalement ces deux suites sont des prédicats libres, présentant la forme suivante :

Cependant, répétons que ces règles ne peuvent pas être appliquées automatiquement, car nous avons vérifié le même phénomène que M. LE GUERN traite dans son article publié dans la revue Le Français Moderne, N + A + SP où l'adjectif peut être rattaché soit au N, soit au SP. Certains adjectifs demandent un complément. C'est pourquoi nous sommes obligés de traiter d'abord les suites prépositionnelles.

Exemple :

- *os papeis dos homens atuantes na organização* (les rôles des hommes intervenant dans l'organisation) D' + N + P' + D' + N + A + P + D' + N Etant donné que N'' □ D' + N, cette description peut être réécrite comme suit : D' + N + P' + D' + N + A + P + N'' Sachant que SP □ P' + N'', la description peut être réécrite comme suit : D' + N + P' + D' + N + A + SP La prochaine réduction serait N □ N + A, cependant nous ne pouvons pas la faire sans savoir ce qui le suit, car le A exige ici un complément. La solution est déjà montrée dans le chapitre 8, lorsque nous avons décidé de créer la variable NC pour informer l'analyseur sur le nombre de compléments d'un mot, qu'il s'agisse d'un verbe, ou d'un substantif ou encore d'un adjectif. Nous voyons qu'il s'agit d'un cas pareil à celui présenté par M. LE GUERN dans son article, de la revue Le Français Moderne, page 31 (« étudiant assidu aux cours »). Ainsi, nous avons adopté la même solution pour le portugais, en le considérant comme étant un prédicat adjectif lié : A : := <prédicat adjectif libre> A' : := <prédicat adjectif lié> A' □ A + SP Cependant, si l'adjectif n'exige pas de complément {ex. : *os recursos financeiros da*

organização (les ressources financières de l'organisation)}, nous pouvons remplacer la suite N + A par N. Cependant, en cas de prédicats adjectifs sans compléments, on ne peut pas faire simplement A' □ A. D'après M. LE GUERN, ***« elle ne convient qu'aux adjectifs qui dominent un N" interne, ceux qu'on appelle parfois les adjectifs de relation par opposition aux adjectifs de qualité… »*** [109]. Nous avons dont la règle suivant pour les prédicats adjectifs liés : A' □ A_{REL} | A + SP N' □ N + A' N □ N + A_{QUA} Après ces considérations, nous pouvons réduire la description de l'exemple donné à : D' + N + P + D' + N + A' En appliquant la règle N' □N' + A', la description est réécrite comme suit : D' + N + P + D' + N' Sachant que N'' □D' + N', la description sera réécrite comme suit : D' + N + P' + N'' En appliquant SP □P' + N'', la description devient : D' + N + SP Finalement, sachant que N' □N + SP, nous pouvons la réécrire comme suit : D' + N', tandis que nous obtiendrons des résultats intermédiaires légèrement différents de celui-ci lorsque nous avons une suite N + A dont l'adjectif est un prédicat adjectif libre, donc sans aucun complément. Par exemple :

– *os recursos financeiros da organização* (les ressources financières dans l'organisation) dont la description est : D' + N + A + P + D' + N En appliquant N'' □ D' + N, la description est réécrite comme suit : D' + N + A + P + N'' En appliquant N □N + A, puisque le prédicat adjectif ne demande aucun complément, la description est réécrite comme suit : D' + N + P + N'' Sachant que SP □ P' + N'', la description est réécrite comme suit : D' + N + SP En appliquant la règle N' □ N + SP, la description est réécrite comme suit : D' + N' Cet exemple montre un prédicat adjectif libre où le complément appartient, en fait, au substantif (« ressource »).

Il faut remarquer aussi que nous avons trouvé, dans le corpus des SN, des prédicats adjectifs libres qui avaient comme complément, non un SP mais un EP. Par exemple :

· as informações consideradas de peso para a instituição (les informations considérées de poids pour l'institution)
· La réécriture de ce SN est : D' + N + A + P' + N + P' + D' + N
· En appliquant N' □ N, la description est réécrite comme suit : D' + N + A + P' + N + P' + D' + N'
· En appliquant N'' □ D' + N', la description est réécrite comme suit : D' + N + A + P' + N + P' + N''
· En appliquant SP □P' + N'', la description est réécrite comme suit : D' + N + A + P' + N + SP
· EN appliquant EP □ P' + N, la description est réécrite comme suit : D' + N + A + EP + SP

L'adjectif « *considérées* » a comme complément l'EP « de poids », en fait, l'EP est liée au prédicat adjectif libre A et non au prédicat nominal N. Ainsi, de manière analogue, nous pouvons établir la règle :

[109] Michel LE GUERN, « Un analyseur morpho-syntaxique pour l'indexation automatique », *Le Français Moderne*, juillet 1991, t. LIX, n°. 3, p. 31.

- A □ A + EP
- Donc, en appliquant cette règle, la description est réécrite comme suit : D' + N + A + SP
- Finalement, sachant que N + A est aussi un prédicat nominal libre N, nous pouvons réécrire cette description comme suit : D' + N + SP
- Ce qui donne comme résultat final : D' + N'

Le prédicat adjectif reste libre car il ne domine pas un syntagme nominal, étant donné que l'EP ne domine qu'un N. C'est pourquoi nous avons adopté cette règle dans la procédure de réduction de règles.

Concernant encore le problème de traitement de la suite de N + A, nous avons retrouvé des SN dont le prédicat adjectif ne concernait pas le N plus proche, mais le N précédent. Voici deux exemples :

as unidades de informação brasileiras (les unités d'information brésiliennes) 1.

Centros de informação brasileiros (Centres d'information brésiliens). 2.

Chacun de ces deux exemples montre que l'adjectif ne concerne pas le mot « information » mais le mot qui le précède (« unités » et « centres »). En fait, la solution serait le repérage des accords en nombres des mots dans les deux cas. Dans le deuxième, la possibilité existe de constater l'accord en genre, car l'adjectif « brésiliens » est masculin et il est en accord avec le mot « centre » qui est aussi un mot masculin. Parmi les variables qui caractérisent un mot dans le lexique, il y a déjà les variables GN et NB qui possèdent des valeurs concernant le genre et le nombre respectivement.

Pourtant nous pourrions adopter une autre solution qui serait de prendre ces expressions « unité d'information », « centre d'information » etc. comme des mots composés. Or, cette mesure surchargerait certainement le lexique. Il y a toujours le souci de ne pas excessivement surcharger le lexique. Tant que les variables, définies ou à définir, pour caractériser chaque unité lexicale, peuvent résoudre les problèmes, nous éviterons de trop charger le lexique.

Or, les deux variables que nous avons indiquées comme de possibles solutions au problème ne sont pas suffisantes, car il y aura des cas où l'adjectif s'accorde aussi bien au nom précédent qu'à l'autre qui précède ce dernier nom.

Exemple : *o estado do avião constante em cruzeiro* (L'état de l'avion constant en croisière)

Cette construction n'est pas usuelle, car la manière la plus courante d'exprimer l'idée de ce SN est : *O estado constante do avião em cruzeiro* (L'état constant de l'avion en croisière.). Pourtant, nous l'avons trouvée dans le corpus. Il faut dire aussi que « *estado do avião* » (état de l'avion) n'est pas un mot composé. Ici, dans ce SN, l'adjectif *constante* se rapporte à l'état de l'avion et non pas à l'avion. Là, le mot *estado* est un FNOM,SIN,MAS, le mot *avião* est un $F_{NOM,SIN,MAS'}$ et le mot *constante* est un FADJ,SIN,GRN. Nous voyons donc que les deux variables GR et NB ne sont pas suffisants pour décider à quel nom se rapporte l'adjectif, car les trois mots sont au

singulier et l'adjectif est non marqué en genre. Il fallait trouver un autre trait sémantique général, par exemple l'indication de niveau ou degré. Les mots *état* et *constant* indiquent le niveau ou degré du mouvement d'un objet. Cependant, nous n'avons pas créé ce type de variable dans la grammaire de référence décrite au chapitre 8, car l'exemple pris était le seul existant dans le corpus et de plus dans une forme peu usuelle, voire inacceptable.

3.3.5 Réécriture des SN avec la suite N + N

Nous avons trouvé dans le corpus, outre les SN avec des pronoms relatifs et de conjonctions de subordination, des SN avec une réécriture différente de celle d'un SN normal. Ce sont les SN suivants :

1. o projeto reconhecimento das margens costeiras do Brasil (Le projet reconnaissance des bords de mer du Brésil) ;
2. tecnologia DBMS(la téchnologie DBMS) ;
3. o chamado Brasil moderno (la dénomination Brésil moderne) ;
4. as entidades do tipo A (les entités du type A)
5. os serviços de informação do tipo bibliotecas (les services d'information du type bibliothèques).

Ce sont des SN avec des configurations descriptives pareilles à la configuration de l'exemple, couramment discutés dans la langue française, lorsque le sujet est SN : Le président Reagan. Ici, d'après M. LE GUERN, Reagan joue, non le rôle d'un nom propre, mais celui d'un prédicat appellatif. Il nous semble que le problème est le même à un détail près : le rôle d'un prédicat appellatif est joué non seulement par des noms propres mais aussi par des lettres, des phrases, des sigles ou de simples ensembles de mots. Ce genre de problèmes n'est pas un privilège des textes en langue portugaise, mais ils peuvent apparaître aussi dans des textes en langue française.

Ainsi, il y a deux problèmes à résoudre : d'abord il faut trouver une solution, une réécriture qui puisse reconnaître ce genre de syntagmes nominaux ; le deuxième problème c'est comment extraire à partir de cette réécriture aussi bien le syntagme nominal complet que les prédicats appellatifs. Ces prédicats appellatifs doivent entrer dans la base de données LEXIQUE comme des unités classées dans la catégorie F_{ADJ}. Ces prédicats appellatifs jouent aussi le rôle d'un SN. En effet, dans la forme de surface il joue le rôle de prédicat appellatif, étant donné qu'il masque son vrai rôle de SN dans la structure profonde. Exemple :

Le président Reagan

Ce SN est, en effet, un raccourci d'un syntagme plus complexe du type :

Le président qui s'appelle Reagan

Ou :

Le président qui est appelé Reagan.

Il nous semble donc important de repérer le SN « Le président Reagan », aussi bien

que le SN « Reagan ».

En fait, dans le contexte de la recherche d'information, un utilisateur difficilement va demander une recherche à partir du centre du syntagme nominal *le président Reagan* (président). Vraisemblablement il doit demander la recherche à partir du nom Reagan. De la même façon, personne ne va demander une recherche sur bibliothèques à partir du syntagme nominal *les services d'information du type bibliothèques*, ou peut-être à partir du syntagme nominal *bibliothèques*. Ainsi, il faut reconnaître les deux syntagmes nominaux.

Il nous semble que dans ce contexte, il faut adopter des algorithmes ad-hoc envisageant d'éviter que les règles du modèle subissent des échecs. Ces algorithmes doivent être démarrés par une règle ou paramètre qui doit caractériser les mots qui peuvent faire partie de ce type de syntagme nominal, précédant le prédicat appellatif, exemple : chamado (nommé), chamada (nommée), tipo (type), presidente (président), diretor (directeur), professor (professeur), engenheiro (ingenieur), etc. Ces mots doivent être caractérisés par la variable RG avec une valeur ADN. Cette valeur indiquera à l'analyseur que le mot ou terme qui le suit doit être repéré aussi comme un SN, bien qu'il soit classé comme F_{ADJ}.

Cet adjectif (le prédicat appellatif) peut être un mot, une lettre ou une phrase. Nous savons que normalement quand il s'agit d'une phrase, et parfois des mots ou des lettres, ils peuvent être entourés par des guillemets. Cependant, cela n'arrive pas toujours, c'est pour cela que nous prenons la décision de le mettre dans le lexique, bien qu'on sache que cette décision peut surcharger le lexique. La règle pour ce type de syntagme sera la suivante :

N'' □ D' + N_0 + A

Nous observons dans cette règle la présence du centre de syntagme N_0, il s'agit d'une restriction pour éviter les problèmes de récursivité qui peuvent se produire avec le prédicat libre N.

3.3.6 Réécriture de SN avec double rection

Etant donné que les syntagmes prépositionnels ne sont pas enchaînés quand on a une double rection, en fait, la réécriture de syntagmes nominaux avec double rection est réalisée en explicitant les syntagmes prépositionnels comme une suite, exemple :

- a análise da informação pelos meios convenientes (L'analyse de l'information par les moyens appropriés)
- La réécriture de ce syntagme nominal est : 1. D' + N + P' + D' + N + P' + D' + N + A
- En appliquant la règle N □ N + A, la description (1) est réécrite comme suit : 2. D' + N + P' + D' + N + P' + D' + N
- En appliquant N' □ N, la description (2) est réécrite comme suit : 3. D' + N + P' + D' + N + P' + D' + N'
- En appliquant N'' □ D' + N', la description (3) est réécrite comme suit : 4. D' + N + P' + D' + N + P' + N''

- En appliquant SP □ P' + N'', la description (4) est réécrite comme suit : 5. D' + N + P' + D' + N + SP
- Ici nous ne pouvons pas appliquer N' □ N + SP car le premier N (analyse) peut demander jusqu'à deux compléments, c'est-à-dire il y a là une double rection. Ainsi, les SP ne sont pas emboîtés. 6. D' + N + P' + D' + N + SP
- En appliquant N' □ N 7. D' + N + P' + D' + N' + SP
- En appliquant N'' □ D' + N', la description (7) est réécrite comme suit : 8. D' + N + P' + N'' + SP
- En appliquant SP □ P' + N'', la description (8) est réécrite comme suit : 9. D' + N + SP + SP
- Ce qui est différent d'un syntagme sans double rection, voyons l'exemple suivant : a base de conhecimento dos países do terceiro mundo (la base de connaissance des pays du tiers monde)
- La réécriture de ce syntagme nominal est : D' + N + P' + N + P' + D' + N + P' + D' + A + N
- En appliquant la règle N □A + N, la description est réécrite comme suit : D' + N + P' + N + P' + D' + N + P' + D' + N
- En appliquant N' □N, la description est réécrite comme suit : D' + N + P' + N + P' + D' + N + P' + D' + N'
- En appliquant N'' □D' + N', la description est réécrite comme suit : D' + N + P' + N + P' + D' + N + P' + N''
- En appliquant SP □P' + N'', la description est réécrite comme suit : D' + N + P' + N + P' + D' + N + SP
- Etant donné qu'ici, le premier N (base de connaissances) ne demande qu'un seul complément, nous pouvons appliquer la règle N' □N + SP, et la description est ainsi réécrite : D' + N + P' + N + P' + D' + N'
- En appliquant N'' □D' + N', la description est réécrite comme suit : D' + N + P' + N + P' + N''
- En appliquant SP □P' + N'', la description est réécrite comme suit : D' + N + P' + N + SP
- En appliquant EP □ P' + N, la description est réécrite comme suit : D' + N + EP + SP
- En appliquant N □ N + EP, la description est réécrite comme suit : D' + N + SP

Ici nous arrivons donc au même niveau de la règle que nous sommes arrivés pour la réécriture du SN avec double rection, sachant que nous pouvons continuer à réduire cette description en appliquant la règle N' □ N + SP et après la règle D' + N', nous arrivons ainsi au N'' qui est le SN.

Nous avons vu aussi le problème de l'ordre d'application des règles et aussi l'ordre d'évaluation et de réduction des descriptions de SN. Selon ce que nous avons dit, l'application des règles de réduction devrait obéir à la direction de gauche à droite.

Cependant, lorsqu'on trouve un déterminant, il faut résoudre d'abord ce dernier SN et ainsi de suite. Cette procédure est nécessaire car s'il y a un déterminant après une préposition il est certain qu'il y aura formation d'un syntagme prépositionnel. Ce qui nous a obligé à donner la priorité de traitement aux derniers syntagmes et donc de faire la procédure de réduction de droite à gauche, lorsqu'on trouvait des syntagmes nominaux intermédiaires.

3.3.7 La détermination des centres de syntagmes nominaux de premier niveau

Selon la définition donnée au début de ce chapitre, le centre de syntagme nominal du premier niveau est le nom (N_0) autour duquel se forme le syntagme nominal. Or, nous avons eu besoin de mieux expliciter ce centre de syntagme nominal quand il est un mot composé, plus précisément lorsqu'il a la configuration suivante :

D' + N + P' + N (le système d'information)

En fait, cette réécriture peut être réduite à D' + N, étant donné que P' + N est une EP et que N + EP est un N, donc le N'' □D' + N. Comme nous l'avons déjà signalé, aussi bien le N (système) que le N (information) et le N (système d'information) sont des prédicats libres. Ainsi, nous avons défini que le centre de syntagme nominal serait représenté par les trois N et non seulement par le N (système). Cette prise de décision permettra aux utilisateurs d'accéder à ce syntagme nominal, non seulement par le centre « système », mais également par le centre « information » et le centre « système d'information », ce qui donnera plus de souplesse et de précision à la recherche d'information. L'indexation seule du centre « système » ne permettra pas aux utilisateurs de trouver le syntagme nominal « le système d'information » à partir du mot « information ». Il fallait à l'usager utiliser le centre « système ». Tandis que si nous indexons les trois centres, nous satisferons non seulement les usagers qui cherchent à partir du centre « système », mais également à ceux qui utilisent le centre « information » et aussi « système d'information ».

Dans la prochaine section nous allons discuter les descriptions de SN les moins courantes, lesquelles sont extraites d'un tableau construit à partir de la consolidation de descriptions.

3.4 Consolidation de la procédure d'établissement de règles

Tout d'abord, selon la méthodologie adoptée, nous avons cherché de trouver des descriptions stables — c'est-à-dire les plus fréquents — qui puissent servir de règle de réécriture des SN, en langue portugaise.

L'établissement des règles a commencé à partir de la description de chaque syntagme nominal, d'abord une simple description traduite en termes de catégories de ses composants. Ces descriptions ont été transformées en règles de réécriture pour le regroupement de tous les éléments du vocabulaire terminal en des éléments du vocabulaire non terminal. Le premier essai de description des syntagmes nominaux nous a fourni 2587 descriptions différentes. Puis, avec le regroupement successif des éléments dans ceux du vocabulaire non terminal, ce nombre a baissé à 323 descriptions

différentes. Ce regroupement a été fait par le biais de règles de réduction.

Or, au fur et à mesure que nous cherchions à réduire le nombre de descriptions, que nous avions supposé être des règles de réécriture des SN, nous nous sommes rendu compte que le nombre de descriptions diminuait. Tandis que la fréquence de certaines descriptions augmentait. Ceci est logique puisque la façon de diminuer les règles a été faite en réécrivant les éléments adjectivaux, les éléments nominaux selon les compositions des unités lexicales. Cette réécriture ne faisait que regrouper ou redéfinir les éléments originaux des SN en ceux du vocabulaire non terminal (N'', N', N, A', A, SP, EP). Le résultat de cette procédure peut être vu dans la figure 10.2, où on trouve un extrait du tableau consolidé de descriptions de SN après la procédure de réduction de descriptions (voir le tableau complet dans l'Annexe C).

Il est clair qu'en adoptant cette démarche, soit nous arrivions à la règle générale de SN — celle de la description numéro 1 du tableau de la figure 10.2, (D' + N') $_{N''}$, —, soit le résultat était composé par elle et aussi par quelques descriptions non réductibles. Dans tous les deux cas, nous sommes arrivés à la conclusion que les règles utilisées pour réduire l'ensemble de descriptions composent aussi la grammaire de reconnaissance et d'extraction de SN. Ainsi, du tableau de la figure 10.2, nous pouvons extraire les descriptions qui peuvent être encore réduites, ce que nous voyons dans la figure 10.3.

Dans le tableau de la figure 10.3, il est facile de voir que presque toutes les descriptions de SN peuvent encore être réduites à la description de numéro 1. Parmi celles qui ne le peuvent pas, se trouvent les descriptions des numéros 14, 19 et 25. Ce que nous allons discuter.

3.4.1 Discussion de la description de numéro 14 : (D' + N + C + D' + N + SP)N''

C'est une coordination entre deux SN cachés pour la coordination de deux compositions du type D + N. C'est-à-dire deux suites de : article + nom. Les syntagmes nominaux décrits pour cette description sont :

- *a crítica e a contextualização do saber* (la critique et la contextualisation du savoir) *Nous voyons là, en effet, qu'il s'agit d'un syntagme qui possède deux syntagmes nominaux cachés derrière la coordination d'une composition de D + N, c'est-à-dire d'un article plus un nom. Les deux syntagmes sont : 1) la critique du savoir ; et 2) la contextualisation du savoir.*
- *a educação e o treinamento para a qualidade* (l'education et l'entraînement pour la qualité) *C'est le même cas montré plus haut, là aussi il s'agit d'un syntagme qu'est à l'origine de deux autres syntagmes nominaux qui sont : 1) l'education pour la qualité ; et 2) l'entrainement pour la qualité)*.
- *a globalização e a integração de grandes grupos econômicos* (la mondialisation et l'intégration de grands groupes économiques) *Là aussi, nous avons un autre syntagme avec les mêmes caractéristiques que les précédents, d'où nous pouvons extraire les syntagmes suivants : 1) la mondialisation de grands groupes*

économiques ; et 2) l'intégration des grands groupes économiques.

- *a importância e a estrutura do fluxo de informação* (l'importance et la structure du flux d'information) *Idem les cas précédents, les syntagmes nominaux existants dans ce SN sont : 1) l'importance du flux d'information ; 2) la structure du flux d'information*.
- *a natureza e o nivel cognitivo da necessidade de informação identificada* (la nature et le niveau cognitif du besoin d'information identifiée) *Là aussi, la suite coordonnée permet d'extraire deux syntagmes nominaux de deuxième niveau comme suit : 1) la nature du besoin d'information identifiée ; 2) le niveau cognitif du besoin d'information identifiée*.

Selon les grammairiens Celso CUNHA & Lindley CINTRA, ***"Lorsque l'article est employé avant le premier nom d'une série coordonnée, l'article doit précéder les noms suivants, encore que soient tous du même genre et du même nombre."*** Et pourtant, selon eux, les articles peuvent être omis lorsque les mots indiquent le même être ou la même chose. L'ensemble de règles que nous avons établi reconnaît les séries coordonnées de noms (N □ N + C + N), comme celles des adjectifs (A □ A + C + A). Or, cet ensemble ne reconnaît pas les constructions de syntagmes nominaux que nous venons de montrer plus haut. Les solutions possibles en ce cas sont : 1) faire un pré-traitement pour supprimer les articles qui précédent les mots de la série coordonnée, à partir du deuxième mot ; 2) établir une nouvelle règle spécifique pour régler ce problème.

L'avantage de la première solution proposée est de ne pas créer une nouvelle règle car les suites de noms ou d'adjectifs sont déjà résolus pour l'ensemble de règles déjà établies. La deuxième solution passe par la création non d'une simple règle mais d'éventuellement de deux règles, comme suit :

SC
::= <série coordonnée de la composition déterminant plus nom>

SC
□ D' + N + C + D' + N | D' + N + <,> + D' + N

N''
□ SC + SP | SC + EP

Il nous semble que la première solution est plus intéressante car elle n'implique pas de créer de nouvelles règles, mais plus simplement de faire un pré-traitement. La création de nouvelles règles dans la création de nouveaux algorithmes pour l'analyse, ceux qui peuvent être compliqué. Tandis qu'en utilisant un pré-traitement nous allons profiter des algorithmes plus simples et aussi de règles qui sont déjà utilisées pour l'analyse de suites coordonnées de noms ou d'adjectifs. Ce qui nous semble plus simple. Nous avons donc deux solutions possibles, la deuxième peut être utilisée si le pré-traitement ne marche pas bien.

3.4.2 Discussion de la description de numéro 19 : $(D' + N + A' + SP)_{N''}$

Cette description porte sur 9 syntagmes nominaux. Voyons d'abord les exemples de SN

qui sont décrits comme cela.

Exemples :

- *as atividades industriais das grandes indústrias* (les activités industrielles des grandes industries)*

  – *industriais* (industrielles) Ce mot peut jouer le rôle d'un adjectif de relation. Ce mot peut dont être réécrit comme suit : *da indústria* (de l'industrie) nous pouvons décrire cette construction comme étant un syntagme prépositionnel comme suit : SP □ P' + D' + N Ici le problème peut être une distorsion causée par la résolution d'un élément anaphorique. Le SN originel était : *ses activités industrielles* où la source de l'élément anaphorique 'ses' était *les grandes industries*. Bien que dans une industrie, il y a des activités non industrielles nous croyons que dans un langage courant on dirait plutôt les activités spécifiques à l'industrie.

- *o aumento populacional do mundo*(l'augmentation démographique du monde)

  – *populacional* (démographique) Nous pouvons adopter le même raisonnement que nous adoptons pour *indústrias*. C'est-à-dire, le mot *populacional* peut jouer le rôle d'un adjectif de relation. Il peut dont être réécrit comme un syntagme prépositionnel.

Or, bien que la forme de surface de ces exemples, puisse faire passer les adjectifs présentés comme étant des adjectifs de relation, ils ne le sont pas. En effet, ces adjectifs dans les contextes présentés peuvent être réécrits comme suit :

- ***atividades industriais*** *das grandes indústrias* (activités industrielles des grandes industries)

  – activités industrielles sont réécrites plutôt comme activités d'industries qu'activités des industries ;

- ***aumento populacional*** *do mundo* (augmentation demographique du monde)

  – *aumento populacional* est réécrit plutôt comme *aumento de população* que *aumento da população*

Il s'agit plutôt de prédicats d'adjectifs libres que des prédicats d'adjectifs liés. En effet, les adjectifs de relation ne jouent pas toujours ce rôle là (relation), parfois ils sont utilisés dans des contextes où ils jouent le rôle de simples adjectifs de qualité. C'est une caractéristique difficile à repérer. M. LE GUERN nos signale qu'une des marques possible est la présence d'un syntagme prépositionnel après l'adjectif. C'est-à-dire que lorsqu'il y a un syntagme prépositionnel après un adjectif qui ressemble à un adjectif de relation, celui ci ne joue plus le rôle d'adjectif de relation mais de qualité. Cette remarque a bien marché pour les 9 cas que nous avons trouvé dans le corpus de SN. C'est-à-dire que la règle

initialement proposée n'existe pas, car le SP est un signe de que l'adjectif précédent est en fait un adjectif de qualité et non pas de relation.

Remarquons qu'une autre description pareille à celle discutée ci-dessus, mais avec la seule différence qu'au lieu d'un SP nous avons trouvé une EP, **(D' + N + A' + EP)** $_{N''}$. Pourtant, elle ne décrit que 3 SN. C'est pour cela qu'elle n'apparaît pas dans le tableau de la figure 10.2 et ni dans le tableau de la figure 10.3. Voyons d'abord les exemples de SN trouvés avec cette description. Exemples :

- *teses norte-americanas de controle de qualidade total* (les thèses nord-américaines de contrôle de qualité totale) ;

  – *norte-americanas* (nord-américaines) [A'] *dos Estados Unidos da Amér*ica (des Etats-Unis) [SP]

- *os fornecedores internacionais de tecnologia* (les fournisseurs internationaux de technologie) ;

  – *internacionais* (internationaux) [A_0] *do estrangeiro* (de l'étranger) [SP]

- *proteção internacional de propriedade* (la protection internationale de propriété).

  – *internacioinal* (International) [A_0] *do estrange*iro (de l'étranger) [SP]

De manière opposée au cas précédent, les exemples présentés plus haut sont des cas où les adjectifs jouent le rôle d'adjectif de relation. Dans ce cas nous proposons la règle suivante :

N' □ N + A_0 + EP

3.4.3 Discussion de la description de numéro 25 : (D' + N + SP + A + SP)$_{N''}$

Cette description n'apparaît que pour décrire sept SN. Avant de procéder à la discussion voyons quelques exemples de SN avec cette description pour connaître quel est le type d'adjectif qui apparaît dans la description.

Exemples :

- *os profissionais da informação atuantes na organização* (les professionnels de l'information actants dans l'organisation)
- *a nova divisão do trabalho inerente ao capitalismo* (la nouvelle division du travail lié au capitalisme)
- *as palavras de Thompson referidas à organização* (les mots de Thompson concernant à l'organisation)

Ces trois exemples sont à l'origine de la description cible dans cette section, c'est-à-dire (D' + N + SP + A + SP)$_{N''}$. Dans les 7 SN, les adjectifs demandent un complément. Ainsi,

le A + SP, doit être pris comme A'.

Ainsi, si nous essayons de réduire cette description, nous avons la description suivante :

- D' + N + SP + A'
- En appliquant la règle N' □ N + SP, la description est réécrit comme : D' + N' + A'

Pour réduire cette description à la règle générale d'un SN, il faut établie une nouvelle règle du type : N' □ N' + A + SP. Cette règle est plus acceptable que la suivante : N' □ N' + A'. Etant donné que A' □ A_0 | A + SP, l'adoption de la règle N' qui réécrit N' + A' peut faire échouer l'analyse parce que A' peut être aussi un adjectif de relation.

3.4.4 Discussion des descriptions avec les catégories Q et V

Les descriptions avec les catégories Q et V sont moins fréquent que les descriptions composées par des unités qui appartiennent à d'autres catégories. En fait, s'il y avait des règles avec des composants comme les pronoms relatifs, les SP ou EP avec des verbes à l'infinitif, et avec des combinaisons de verbes auxiliaires et au participe, nous pourrions vraisemblablement arriver à réduire toutes les règles (descriptions des SN) à une seule, c'est-à-dire la règle générale du SN. Le tableau de la figure 10.4 le montre.

Examinons quelques-unes unes de ces descriptions :

8. $(D' + N + Q_{\text{-que}} + V + (D' + N')_{N''})_{N''}$
où Q : := <unité de la catégorie Q, où seront placés les pronoms relatifs>

Exemples de SN avec cette description :

- *a premissa que acompanha os servicos de consultoria* (la prémisse qui accompagne les services de consultation)
- *a realidade que vive a organização* (la réalité que vit l'organisation)
- *o conhecimento que gerou a invenção* (la connaissance qui a généré l'invention)

Cette description décrit 32 syntagmes nominaux. Etant donné la décision prise de ne pas traiter les SN avec pronoms relatifs, puisque nous n'avons pas bien caractérisé la catégorie Q du lexique nous laisserons ce type de descriptions et toutes celles qui impliquent l'utilisation des pronoms relatifs et aussi des conjonctions de subordination pour une recherche à accomplir dans l'avenir.

Finalement, d'autres descriptions moins importantes sont celles qui impliquent l'utilisation de verbes, exemple :

12. $(D' + N + P' + V + (D' + N')_{N''})_{N''}$(cette description décrit 16 SN)

- *a capacidade de manipular a informação* (la capacité de manipuler l'information) *Ce SN peut être aussi exprimé comme : la capacité de manipulation de l'information*.
- *a capacidade de manusear informações* (la capacité de manier des informations) *Ce SN peut être aussi exprimé comme : la capacité de maniement des informations*

- *a estratégia de alcancar a auto-sustentação* (la stratégie pour atteindre l'auto-sustentation) *Ce SN peut être aussi exprimé comme : la stratégie d'atteinte de l'auto-sustentation*.
- *a função de planejar a informatização* (la fonction de planifier l'informatisation) *Ce SN peu être aussi exprimé comme : la fonction de planification de l'informatisation.*

Nous avons montré quelques exemples de SN qui ont été décrits par le numéro 12, du tableau de la figure 10.3, avec des verbes à l'infinitif après une préposition. Les seize SN ont la même configuration. Nous avons vu aussi que ces SN peuvent être modifiés en utilisant un nom N (nominalisation d'un verbe) avec un SP. Pourtant, nous allons laisser ce genre de SN pour une autre recherche, afin de mieux comprendre l'utilisation et le comportement des verbes dans les SN.

26.$(D' + N + P' + V + (D' + N + SP)_{N''})_{N''}$(cette description décrit 7 SN)

- *a preocupação de manter o equilibrio entre dois grandes objetivos* (le souci de maintenir l'équilibre entre deux grands objectifs) *Là, la différence est l'existence d'un SN de deuxième niveau après le verbe. Mais d'une manière générale, la configuration est pareille à la description antérieure. Nous pouvons remplacer le verbe maintenir par conservation ou maintient*.
- *capacidade de conhecer as tecnologias da informação* (la capacité de connaître les technologies de l'information)
- *o objetivo de mostrar tendências de tecnologias* (l'objectif de montrer les tendances des technologies)
- *o objetivo de fomentar a adoção de inovações tecnológicas* (l'objectif d'encourager l'adoption des innovations technologiques) *Les observations faites pour le premier exemple sont aussi valables pour ces trois derniers exemples*.

Le but de montrer ces descriptions, considérées ici comme moins importantes, est premièrement pour dire qu'elles existent et deuxièmement pour expliquer que bien qu'elles soient simples et passibles d'être analysées, nous ne les avons pas faites. Ce type de SN a une grande variation de composition. Il faut comprendre ces variations et chercher des règles qui puissent réduire ces descriptions à la règle générale d'un SN.

Nous ne pouvons pas tout résoudre actuellement, il faut d'abord établir une grammaire pour les SN les plus courants et après, dans un contexte déjà connu, dans une phase de perfectionnement, traiter les SN qui sont composés par des éléments de la catégorie Q et V.

Dans la prochaine section nous allons examiner d'autres types de descriptions encore moins fréquents.

3.4.5 Examen de quelques descriptions les moins fréquentes

Il est vrai qu'il y aura toujours des descriptions de SN sui generis, puisqu'on prend, par définition, les titres de documents et de sections comme des SN. Ce fait pourra empêcher la réduction totale de l'ensemble de descriptions des SN du corpus à la règle générale.

Nous allons examiner les descriptions de SN les moins courantes et qui ont des règles de réécritures singulières. Ces descriptions n'apparaissent pas dans le tableau de la figure 10.2 puisqu'elles ont une fréquence d'occurrence très faible, elles n'apparaissent que dans le tableau de l'annexe C. Dans les prochaines sections nous allons discuter la réécriture de titres d'articles (et de sections), et aussi d'autres descriptions qui éventuellement peuvent aider à la formation de futures règles.

3.4.5.1 Réécriture de SN de titres

Dans cette section nous n'allons pas proposer une règle de réécriture de SN du type titre, mais discuter les problèmes de ce type de SN, étant donné que leurs règles de réécriture très sont variées. Cette variation est due au fait qu'ils sont pris comme des SN par définition et non à cause de leur syntaxe.

Les titres d'articles et titres de sections d'un article n'obéissent pas toujours aux règles de réécriture des SN trouvés dans les textes. En portugais, comme peut-être dans plusieurs autres langues, on trouve des titres avec deux points, exemple :

- Interação entre empresas com necessidades de informação (=conhecimento) e a estrutura nacional de centros com provisão de conhecimento acumulado: referência especial à estrutura nacional de serviços de informação, documentação e de biblioteca ;
- Gerência da Informação : mudanças nos perfis profissionais
- Informação : instrumento de dominação e de submissão
- Informação: a chave para a qualidade total
- Informação Técnico-econômica : mais importante do que nunca
- Sistemas de Informação : a evolução dos enfoques
- Consultoria Informatológica em revisão : uma alternativa para serviços de informação personalizados

La ponctuation « deux points » est présente dans presque la moitié des titres des articles du corpus (15 articles au total). Les titres (1) et (7) sont respectivement des articles traduits de l'anglais et de l'espagnol. Outre cette caractéristique remarquée, il est possible aussi de voir l'absence d'article devant les titres, comme le remarquait M. LE GUERN dans son article [110] de la revue *Opérateurs et constructions syntaxiques : Evolutions des marques et des distributions du XV ème au XX ème siècle*. D'ailleurs, nous voyons là aussi l'apparition du signe de ponctuation deux points, tant dans le titre de la revue que dans le titre de l'article.

Parmi les exemples cités, le premier (1) nous semble le plus curieux des titres présentés, non seulement à cause des deux points, mais à cause aussi de sa taille et de la présence d'un terme entre parenthèses avec un signe d'égalité. Les parenthèses et le

[110] Michel LE GUERN. « Traitement automatique et variation linguistique : la syntaxe des titres ». *Opérateurs et constructions syntaxiques : Evolutions des marques et des ditributions du XV ème au XX ème siècle.* Paris :Presses de l'Ecole Normale Supérieure, 1994, p. 75-81.

signe d'égalité sont inattendus dans un titre, puisque d'une manière générale on cherche à mettre une phrase courte et le plus simple possible.

En ce qui concerne l'absence d'article, nous pouvons suivre la solution proposée par M. LE GUERN, en mettant l'article défini devant le titre.

Par rapport au problème de l'existence du signe de ponctuation deux points, nous proposons de faire l'extraction des SN dans la partie qui précède les deux points, et dans la partie qui les suit, de manière indépendante. Cette proposition vient du fait que la deuxième partie d'une phrase avec « deux points » est une forme d'explication de la première partie qui se trouve avant ces « deux points ».

On peut trouver aussi la deuxième partie de ces phrases comme étant des compléments à la première partie, c'est-à-dire, celle qui est avant les deux points. Or, nous n'avons pas trouvé des titres avec cette particularité. Cependant, il peut apparaître dans les textes. Nous n'avons pas trouvé cela dans le corpus de SN. Cette absence est due au fait que l'extraction a été faite de manière artisanale, c'est-à-dire, par nous-même et nous avons déjà sûrement réglé ce type de problème au moment de l'extraction.

Le traitement automatique du problème des compléments des SN qui se trouvent après les deux points passe par l'analyse de la première partie qui précède ceux-ci. Il faut vérifier si le terme qui vient juste avant les deux points est une préposition, ou si le mot qui les précède demande un complément. Si un de ces deux cas se présente, il faut combiner cette première partie avec celle qui suit les deux points. Pour faire la combinaison il faut analyser la deuxième partie aussi, en cherchant à savoir s'il s'agit d'une suite coordonnée, soit de mots ou de syntagmes prépositionnels ou même d'expansions prépositionnelles. Dans le cas de suites coordonnées, il faut faire la factorisation ou ce que Omar LAROUK appelle dans sa thèse de « calculs des images logico-semantiques ». C'est-à-dire, faire la combinaison distributive de la première partie avec chacun des termes qui se trouve dans la suite coordonnée de la partie qui suit les deux points.

En bref, la solution proposée pour la reconnaissance de SN dans les titres est très simple. Il faut d'abord, s'il n'y a pas d'article, mettre un article défini au début du titre et en suite chercher à y reconnaître les SN comme s'il s'agissait d'un texte quelconque, en utilisant les règles de réécriture des SN, et en tenant compte bien sûr des problèmes de l'existence de ponctuation.

3.4.5.2 Descriptions de faux SN

Nous avons trouvé encore des faux SN, en prenant comme exemple une de ces descriptions dues au repérage de faux SN :

D' + W + N

Outre le fait que ce sont des faux SN, ils ont été mal décrits car nous avons commis une erreur dans la procédure de description des SN, comme nous pouvons le constater :

- os mais baratos (les moins chers)
- os mais eficientes (les plus efficaces)
- *pouco produtivo (peu productif*)

- *suficientemente inteligente (suffisamment intelligent) Dans ces quatre SN nous avons décrit comme N des mots qui sont en fait des adjectifs. De plus, dans les deux premiers SN (Os mais baratos e os mais eficientes), il y a un élément anaphorique qui n'a pas été repéré*.
- pouco valor (peu de valeur)
- mais recursos (plus de ressources) Ces deux derniers SN ont été décrits correctement, mais il manque quelque chose car ils ne donnent pas un sens précis.

Il est vrai que nous avons trouvé quelques descriptions de SN mal faites, à la dernière minute, dans la phase de rédaction de cette thèse, certaines erreurs ont été trouvées comme dans les exemples montrés. Pourtant, nous avons revu le corpus de SN maintes fois, plusieurs corrections et ajustements ont été faits dans le corpus de manière à éviter ces problèmes. Mais c'est une des tâches la plus difficile à cause du volume de SN dans le corpus. Bien que nous ayons relu plusieurs fois attentivement il arrive que des erreurs passent sous nos yeux sans et échappent au regard. C'est peut-être paradoxalement à force de relire, corriger et d'analyser ce corpus plusieurs fois.

3.5 Consolidation de la grammaire de reconnaissance et d'extraction de SN

Etant donné que les règles ont été dispersées dans ce chapitre lors de la discussion, nous allons maintenant les regrouper pour les consolider dans une grammaire de reconnaissance et d'extraction de SN.

- Les déterminants complexes

D' □ D0 | D | E | E + DNUM | DDEF + E + DNUM | WQUA + DNUM | WQUA + P-DE +1. DNUM | DIND + I |

I □ E_{INT} - E_{INT} | E_{INT}

D0
: := <déterminant zéro (l'absence de déterminant)>

D
: := <unité de la catégorie des D>

E
: := <unité de la catégorie D, sous-catégorie nombres - E>

EINT
: := <unité de la catégorie D, sous-catégorie nombres E, entiers>

EDEC
: := <unité de la catégorie D, sous-catégorie nombres E, décimales>

DNUM
: := <unité de la catégorie D, sous-catégorie numériques>

DNNU
: := <unité de la catégorie D, sous-catégorie non numériques>

DIND
: := <unité de la catégorie D, sous-catégorie indéfinis>

I
::= <fourchette numérique>

WQUA
: := <unité de la catégorie des Adverbes (W) de quantité>

D □ DNUM | DNNU | DDEF | DIND 1.

E □ EINT | EDEC 2.

· Les nominaux

N0□ FNOM | FNAN 1.

N □ N0 | N + A | A + N | N + EP | N + C + N 2.

N0
::= <centre du syntagme nominal>

FNOM
: := <unité de la catégorie F, nom>

FNAN
: := <unité de la catégorie F, non marqué>

FNOM,PRP
: := <unité de la catégorie F, noms propres>

N
: := <prédicat libre>

N'
: := <prédicat lié>

N''
: := <syntagme nominal>

EP
: := <expansion prépositionnelle>

SP
: := <syntagme prépositionnel>

C
: := <élément de la catégorie conjonctions de coordination (e/ou)>

N' □ N + A' | N' + C + N' | N + SP | N' + EP | N' + SP | N0 + A | N + A' + EP | N' + A + SP | N 1.

N'' □ D' + N' | FNOM, PRP | FNOM, PRO | D' + N + SP + SP | D' + N + SP + SP + SP2. | D' + N + SP + SP + SP + SP* Par définition, les noms propres (FNOM,PRP) et les noms pronoms (FNOM,PRO) ont été pris comme des syntagmes nominaux, ainsi que les titres d'articles et de section. Comme noms propres nous avons considéré aussi

les sigles et les noms d'entreprise. * En ce qui concerne les doubles rections, le professeur Emilio GIUSTI, professeur de linguistique portugaise à l'Université Lumière – Lyon 2, nous a rappelé que nous pouvons rarement rencontrer des mots qui demandent jusqu'à 4 compléments. Bien que nous ayons trouvé que des mots qui demandent deux compléments, nous avons décidé d'en prévoir quatre compléments possibles.

· Les adjectivaux

A □ FADJ, QUA | FNAN,QUA | A + C + A | A + WAAJ | WAAJ + A | A + EP 1.

A' □ A0 | A + SP 2.

FADJ, QUA
: := <unité de la catégorie NA, adjectif de qualité>

FNAN,QUA
: := <unité de la catégorie NA, non marqué, qui peut jouer le rôle d'adjectif de qualité>

· Les Syntagmes prépositionnels

P' □ P | P0 1.

SP □ P' + N'' 2.

P'
: := <préposition complexe>

P
: := <unité de la catégorie P – les prépositions simples>

P0
: := <unité de la catégorie P - les locutions prépositionnelles>

· Les Expansions prépositionnelles

EP □ P' + N 1.

Après la procédure complète de réduction des descriptions, nous sommes arrivés à 323 descriptions. Dans celle-ci 107 décrivent plus d'un SN, tandis que 216 n'en décrivent qu'un. Cela représente 3,61 % des SN du corpus. Nous avons déjà dit que cela s'explique par la grande variation de composition des unités de la catégorie des conjonctions de subordination, les pronoms relatifs y compris, des unités de la catégorie des verbes. Ces descriptions peuvent peut-être, être regroupées au moment où on arrive à définir des règles avec ces unités. Il faut dire que parmi ces descriptions on trouve aussi quelques titres de section puisque dans quelques articles, les auteurs écrivaient dans un style qui privilégiait une écriture par items, c'est-à-dire qu'ils écrivaient des phrases pareilles aux

titres, sans utilisation de déterminants, en mettant les mots, normalement les noms, les uns à côtés des autres, comme ceux qui suivent :

- *atualização – palavra-chave* (mise à jour – mot clé)
- *atualização - prospecção tecnológica* (mise à jour – prospection technologique)
- *atualização - prospecção tecnológica e de inteligência* (mise à jour – prospection technologique et d'intelligence)

Certaines descriptions de SN, parmi ceux de fréquence égale à 1, peuvent être réduites à la règle générale. Ces sont des descriptions de SN d'une longueur importante, normalement les SN du plus haut niveau, comme ceux du cinquième niveau. Ils ne sont donc pas nombreux.

4 Considérations sur l'ordre d'application des règles

La démarche adoptée nous a montré que les règles obéissent à un ordre d'application pour trouver la bonne description réduite. Il est vrai qu'à ce moment là nous faisions des réductions d'une description. C'est-à-dire que nous faisions la synthèse d'une description, le regroupement des éléments composants d'un SN. Tandis que la procédure d'analyse pour chercher à reconnaître et à extraire des SN, doit être l'inverse de la procédure de synthèse d'une description. Mais il faut cependant prendre soin à l'ordre de reconnaissance et d'analyse des mots d'un texte, à la lumière des règles conçues pour la grammaire de reconnaissance et d'extraction de SN.

Par exemple, l'absence de déterminant dans un SN ($D_{\varnothing}$), elle ne peut pas être trouvée ou déterminée avant la fin de la procédure de l'analyse du déterminant, lorsque l'algorithme n'en trouve aucun. En ce cas, on dira qu'il s'agit d'un déterminant zéro, si — et seulement si — l'unité prise est un nom au pluriel ou un nom abstrait. Dans le cas de noms abstraits au singulier on peut trouver des marques d'absence de déterminant si l'unité lexicale précédant n'est pas une préposition "de". Il s'agira alors d'absence de déterminant, ou $D_{\varnothing}$. Nous avons montré dans ce cas que les algorithmes de reconnaissance d'un déterminant zéro ne peuvent être appliqués qu'à la fin de l'analyse de toutes les règles de déterminants. C'est la dernière règle à tester. Le même phénomène arrive lors de la reconnaissance d'un nom simple (N_0 ou N) comme étant un prédicat lié. Un N ne peut être considéré comme un prédicat lié que lorsqu'il n'y a pas d'autres éléments qui appartiennent au SN après lui. C'est-à-dire que ce N pour être pris comme un prédicat lié doit être le dernier élément du SN qu'on est en train d'analyser.

Un autre exemple : l'ordre de règles de réécriture d'un SN change, dans une analyse d'un texte, lorsqu'on trouve un déterminant. C'est alors une marque de début d'un nouveau SN. Il faut d'abord reconnaître ce nouveau SN pour ensuite compléter l'analyse de reconnaissance du SN dont ce SN fait partie. C'est-à-dire qu'un déterminant change l'ordre d'application de règles dans une procédure de reconnaissance de SN.

Nous n'allons pas donner l'ordre exact de chaque règle de la grammaire de reconnaissance et d'extraction de SN dans ce travail, car il est nécessaire de beaucoup réfléchir encore là-dessus. L'intérêt de cette remarque est d'alerter ceux qui veulent

développer ce modèle. Nous ne pouvons pas généraliser la nécessité de faire attention à l'ordre d'application de règles à l'ensemble de celles de la présente grammaire, puisque cette remarque dépend peut-être de l'approche de développement sur l'ordinateur de ce modèle.

En effet, la question d'établir un ordre d'application de règle dans la procédure de reconnaissance des SN concerne plutôt le perfectionnement du modèle, ces questions sont résolues au moment de l'implémentation du modèle sur un ordinateur. Ce sont des ajustements fins de dernier moment. Le but est d'attirer l'attention, il faut prendre soin à ce petit détail lors du développement. Les règles consolidées dans cette section ne sont dans aucun ordre particulier.

5 Conclusion

Cette étape de l'étude a été la plus importante et aussi la plus longue à travailler car il a fallu analyser et procéder à des changements dans les descriptions de SN. L'établissement d'une méthodologie de travail a été très important dans la procédure d'analyse et d'application de règles de réduction. Il a fallu même tester l'application de règles de réduction plusieurs fois sur tout le corpus. Ceci provient de mauvaises décisions prise initialement, comme celle de supposer que tous les mots composés seraient mis dans le lexique. Ainsi le travail a dû être remis plusieurs fois sur l'ouvrage. Au bout du compte, il nous semble que nous sommes arrivés à un bon modèle. La grammaire que nous avons construite peut reconnaître au moins 90% de SN. Nous pensons même qu'avec cette grammaire, nous pouvons arriver à reconnaître un pourcentage de SN plus élevé, puisque nous sommes arrivés à cette valeur, en comptant seulement les descriptions qu'ont une fréquence supérieure à 2. Or, nous avons parmi les descriptions dont la fréquence est inférieure à 2 des descriptions qui peuvent être reconnues par la grammaire que nous venons de construire. Malgré l'existence prédominante, parmi ces descriptions, éléments qui appartiennent à la catégorie Q e V, et aussi d'autres problèmes que nous avons déjà signalés, certaines descriptions ne se compte pas avec ces derniers éléments (Q et V). Ce sont les descriptions des SN de plus haut niveau.

En tenant compte du but de ce travail, il nous semble que le pourcentage de reconnaissance de 90% des SN est satisfaisant. Pour résoudre un peu moins de 10% des descriptions du corpus de SN, l'effort serait tellement important qu'il paraît difficile de l'entreprendre actuellement. En outre, dans les SN qui ne sont pas reconnus par notre grammaire, leur importance, pour l'indexation automatique et la recherche d'information, n'est pas certaine.

Ont-ils un statut de descripteur ? C'est ce qu'il faut savoir avant de chercher à établir une grammaire spécifique qui privilégie les pronoms relatifs, les conjonctions de subordination, les verbes.

La seule mesure que nous avons pour le moment, est que ces SN ne représentent qu'environ 10% de tous les SN extraits du corpus d'articles. Et de plus, une grande partie des configurations de ces SN n'a qu'une fréquence égale à 1. Bien que nous ne puissions pas donner une réponse définitive et sûre, avec ces arguments seulement, il nous semble

que ces nombres montrent surtout la faible occurrence de ces SN par rapport aux SN qui peuvent être reconnus par la notre grammaire.

La vie est l'art de tirer des conclusions suffisantes de prémisses insuffisantes . (Life is the art of drawing sufficient conclusions from insufficient premises.) Butler (Samuel),[1835 - 1902], Notebooks.

Conclusion

1 Présentation

Dans ce travail, dans lequel nous avons touché au moins deux grands domaines — la linguistique et les sciences de l'information —, nous mènent à détailler la conclusion concernant à, au moins, quatre aspects principaux, à savoir :

1. la proposition d'un nouveau système de recherche d'information, ce que nous appelons Système de Recherche d'Information Assistée par Ordinateur ;
2. le modèle pour la reconnaissance et l'extraction des syntagmes nominaux (SN) ;
3. les recherches à accomplir dans l'avenir en vue de compléter le modèle de reconnaissance et d'extraction des SN en langue portugaise et perfectionner l'approche proposée ; et
4. l'avenir des SN dans d'autres applications.

Ainsi, ce chapitre est partagé dans ces quatre aspects.

2 La proposition d'un Système de recherche d'information assistée par ordinateur

Les traits généraux de cette proposition ont été signalés par M. LE GUERN dans son article dans la revue *Le Français Moderne*. Nous avons construit une maquette d'un système de recherche d'information suivant la démarche proposée par lui.

Il est vrai que nous n'avons pas fait une évaluation de cette maquette avec des utilisateurs et, donc, nous ne savons pas l'impact de cet outil dans le milieu des utilisateurs. C'est-à-dire que nous n'avons pas des données sur l'efficacité ou sur la performance de ce système. Nous ne pouvons donc pas affirmer avec certitude que ce système résoudra les problèmes discutés dans la première partie de cette thèse.

Nous n'avons pas fait l'évaluation de la maquette puisque cela n'était pas le but de cette recherche. En effet, un travail d'évaluation d'une maquette d'un système de recherche d'information impliquerait d'en avoir deux éléments principaux : a) un corpus de documents important, dans un domaine bien précis ; et b) un ensemble d'usagers aussi important et homogène.

Le corpus n'avait pas une taille adéquate pour l'évaluation, mais il avait une taille suffisante pour l'expérimentation de la maquette construite. La constitution d'un corpus pour l'évaluation d'un système de recherche d'information devrait avoir une taille plus importante. L'évaluation de la maquette dans cette phase de la recherche demanderait l'utilisation d'un système de reconnaissance et d'extraction automatique de SN. Cela est très important pour le traitement et l'indexation d'un corpus plus grand que celui que nous avons pris. En effet, la tâche de reconnaissance et d'extraction manuelle de SN n'est pas facile. Là, il y a deux problèmes majeurs : 1) le temps nécessaire pour la reconnaissance, l'extraction et l'indexation manuelle de SN est excessivement grand ; 2) la procédure manuelle de repérage des SN ne permet pas leur extraction de façon homogène. Un système de reconnaissance, d'extraction et d'indexation automatique des SN serait essentiel. L'utilisation d'un tel système permettrait sûrement une reconnaissance et une extraction plus homogène des SN car il est systématique et obéit plus précisément aux règles de réécriture des SN. Ce qui n'arrive pas souvent dans une démarche manuelle de reconnaissance des SN. De même, nous avons exploité la maquette en utilisant un thesaurus de sciences de l'information, en langue portugaise. Cela a permis de connaître un peu mieux la maquette de recherche d'information, la démarche de navigation dans la maquette, les problèmes de navigation et de la structure en arbre des SN.

Cette exploitation nous a permis quelques réflexions sur l'approche adoptée pour le développement de la maquette comme proposition d'un système de recherche d'information assistée par ordinateur. D'abord nous allons faire quelques remarques sur les faiblesses de la maquette développée dans cette recherche :

- Le début d'une recherche d'information ne doit pas être limité au centre de syntagme

nominal de premier niveau. Cela est intéressant pour les usagers non spécialisés. Cependant, les usagers plus expérimentés peuvent être gênés s'ils trouvent seulement cette démarche de recherche d'information. Ainsi, il faut prévoir dans le système de recherche d'information proposé la possibilité de commencer la recherche d'information à partir d'un centre de syntagme nominal de plus haut niveau. Cela évitera à ces usagers de perdre du temps en naviguant dans l'arbre des SN dès le début de leur arbre ;

- Il est souhaitable aussi que le nouveau système de recherche d'information puisse offrir la recherche d'information directement à partir d'un syntagme nominal donné. C'est-à-dire, de trouver les documents directement à partir d'un syntagme nominal complet, sans avoir le besoin de naviguer dans l'arbre de SN, étant donné que l'usager sait parfois exactement ce qu'il veut. En d'autres mots, l'usager a déjà le syntagme nominal à l'esprit et il n'a pas besoin de naviguer dans l'arbre de SN ;
- Dans la perspective de donner d'autres choix de navigation dans la base de données, il nous semble important qu'un tel système ait des liens hypertextuels qui puissent lier un document à l'autre. Par exemple, relier tous les documents d'un même auteur, à travers un lien, pourra aider les usagers à construire d'autres démarches de recherche d'information dans la base de données. Les liens hypertextuels, par exemple, dans les champs référentiels (auteur, titre du document, titre de la série, etc.) dans la base de données permettront au système d'offrir aux usagers plus de convivialité. C'est une façon de permettre aux usagers de reformuler leurs stratégies de recherche d'information, par le biais d'une navigation non séquentielle dans l'ensemble des documents d'une base de données.

La principale réflexion que nous pouvons faire, par rapport au système proposé, est sur la nouvelle démarche de recherche d'information qu'il offre aux usagers. Les systèmes classiques de recherche d'information utilisent un seul plan de recherche. C'est-à-dire que la recherche est faite directement dans l'ensemble de documents d'une base de données. Il est vrai que la procédure de traitement de la requête et d'appariement de la requête avec les documents d'une base de données n'est pas faite directement dans les textes des documents eux-mêmes mais en utilisant un fichier d'indices ou des fichiers inversés, qui contiennent les descripteurs ou des indices avec une liste de tous les documents dont ils ont été extraits. La réponse est toujours un ensemble de documents trouvés selon ce qu'il y a dans la requête.

La proposition que nous venons de faire, offre aux usagers une démarche différente. Au lieu d'utiliser un seul plan de recherche d'information, elle y est faite en deux plans : 1) celui de la structure de SN ; 2) celui de l'ensemble de documents. Cela peut être représenté par la figure 11.1.

La figure 11.1 montre le schéma de navigation du système de recherche d'information assisté par ordinateur. La navigation est faite en deux plans, d'abord sur l'arbre des SN et lorsque les usagers trouvent le syntagme nominal qui satisfait leur besoin d'information, ils font l'accès aux documents d'où ce syntagme a été extrait. Selon encore cette proposition, les usagers peuvent naviguer dans le plan de l'ensemble de

documents à partir des liens hypertextuels existant parmi les documents.

L'avantage de cette approche est le fait que ce sont les usagers qui font la recherche, naturellement aidés par l'ordinateur, puisque ce sont eux qui décident les documents qui satisfont le mieux leurs besoins d'informations. C'est là la différence entre cette approche et celle des systèmes classiques de recherche d'information. De manière opposée, les systèmes classiques cherchent et décident eux-mêmes, suivant ce que la requête leur demande, quels sont les documents que satisfont la requête et non pas les besoins d'information des usagers. Dans ce cas-là, on suppose que la requête est capable d'exprimer tout le besoin d'information des usagers. Ce qui n'est pas toujours possible.

Un autre avantage c'est le fait qu'en naviguant sur la structure des SN, les usagers apprennent ce qu'il y a dans la base de données. A part ces avantages, il faut remarquer que dans cette approche les usagers n'utilisent ni un langage de commande et ni des opérateurs booléens pour la formulation de leurs requêtes ou demandes d'information. Il ne faut donc apprendre ni l'utilisation d'un langage artificiel de commande, ni l'utilisation d'opérateurs booléens ni des connaissances de logique booléenne. Bien que cette maquette n'ait pas été soumise à une évaluation, les avantages cités et les caractéristiques interactives du système proposé paraissent pouvoir donner beaucoup plus de convivialité aux usagers que les systèmes classiques.

Ainsi, selon ce que nous avons montré là, nous pouvons consolider notre proposition en faisant un petit bilan, en énumérant ses caractéristiques principales :

- le traitement et l'indexation automatique des documents de la base de données sont faits par le biais de la reconnaissance, l'extraction et l'indexation des SN, en construisant une structure en arbre ;
- les documents sont stockés dans un format en langage SGML, avec des liens hypertextuels localisés dans quelques champs référentiels comme : auteurs, éditeurs, titre de publication, etc. Il est souhaitable aussi de faire des liens avec les syntagmes nominaux existant dans ces documents. Pour cela, il faut avoir des critères d'établissement de ces liens. On ne peut pas faire de liens hypertextuels avec tous les SN extraits ;
- l'interface doit utiliser des facilités graphiques, de la couleur et de la souris pour qu'on puisse avoir plus d'interactivité et de convivialité ;
- les menus sont construits de manière dynamique, c'est-à-dire à partir de la structure arborescente des SN de manière à permettre aux usagers d'y monter ou descendre ;
- Dans tous les écrans, l'interface de recherche d'information doit permettre à l'usager de revenir sur le niveau précédent du syntagme nominal. Il faut d'ailleurs que l'interface permette non seulement cela, mais aussi la possibilité de revenir sur le début de la recherche d'information, soit le premier écran. L'interface doit permettre aussi à l'usager de voir les documents d'où un syntagme nominal choisi a été extrait ;
- L'interface doit être munie d'un système d'aide contextuelle à l'usager ;

Bien que nous ayons utilisé un système de gestion de bases de données relationnelles, il faut pour un système professionnel, construire un système sur mesure avec une structure

de données appropriées à la structure arborescente des SN. L'usage d'un système de gestion de bases de données marche bien pour une maquette ou pour une application administrative, mais pour une application professionnelle de recherche d'information il faut adopter une approche de développement la plus performante possible.

En concluant cette section, il faut remarquer que les structures de données présentées dans le chapitre huit sont appropriées à la navigation dans les syntagmes nominaux, selon les caractéristiques de la maquette développée dans le cadre du DEA. Pourtant, si on veut implémenter les facilités hypertextuelles et d'accès au SN de plus haut niveau à partir de son centre, il faut dessiner des nouvelles structures de données.

3 Le modèle de reconnaissance et d'extraction automatique des SN

Le modèle de reconnaissance et d'extraction automatique des SN a été conçu de manière spécifique pour l'indexation automatique. Les solutions adoptées ont été décidées en fonction de cet objectif. Il est évident que ce modèle n'est pas encore complet. Il y a des tâches à accomplir que nous allons énumérer par la suite. De toute façon le modèle présenté dans cette thèse est capable de reconnaître au moins 90% des syntagmes nominaux dans les textes en langue portugaise, dans le domaine des sciences de l'information.

La démarche suivie dans cette recherche nous a facilité la construction de ce modèle car la procédure de reconnaissance et d'extraction manuelle des syntagmes nominaux a beaucoup aidé la prise de connaissance du comportement des syntagmes nominaux dans les textes en langue portugaise. Il a permis d'apercevoir des problèmes comme l'absence des déterminants dans des nombreux syntagmes nominaux en langue portugaise. Nous avons aperçu de grandes similitudes entre les règles de réécriture des syntagmes nominaux dans les langues française et portugaise. Cela s'explique puisque les deux langues ont comme même origine le latin.

3.1 Les recherches pour accomplir dans l'avenir

Le modèle conçu dans cette thèse n'est pas accompli. Il faut le compléter. Selon ce qu'on a déjà dit, le modèle conçu peut reconnaître un ensemble d'à peu près 90% des syntagmes nominaux d'un texte en langue portugaise. Il faut chercher à arriver aux environs de 100% de reconnaissance des syntagmes nominaux dans ces textes puisqu'il y a d'autres applications dans le champ des sciences de l'information où l'utilisation des syntagmes nominaux est parfaitement appropriée. Nous en parlerons dans la prochaine section. Ces applications sont naturellement différentes de l'indexation automatique, pour cela le modèle actuel peut ne pas être suffisant.

Ainsi nous allons énumérer les points nécessaires pour compléter de ce modèle :

1. En ce qui concerne la grammaire de référence, c'est-à-dire la partie de caractérisation des unités lexicales, il faut étudier et détailler le traitement des pronoms relatifs et des conjonctions de subordination pour avoir une définition complète de la catégorie Q. Cela doit être développé avec l'établissement d'une syntaxe pour la reconnaissance des syntagmes nominaux qui contiennent ces unités ;

2. Il faut aussi étudier de plus près les éléments anaphoriques et proposer une sorte de traitement qui puisse résoudre les sources de ces éléments de manière à faciliter la reconnaissance des syntagmes nominaux cachés dans ces éléments. Parmi ces éléments anaphoriques, il y a ceux qui forment de nouveaux syntagmes nominaux et il y a ceux qui n'en forment pas. La reconnaissance des syntagmes nominaux formés dans le premier cas est importante pour l'indexation automatique puisqu'ils peuvent former des syntagmes nominaux de niveau plus haut. Tandis que les syntagmes qui ne forment pas des nouveaux syntagmes, ceux qui constituent plutôt une nouvelle occurrence, bien qu'ils puissent être écartés de l'indexation automatique, mais ils sont importants pour les études d'analyse de contenu. Il s'agit de ceux qui sont basés sur le comptage des cooccurrences de syntagmes nominaux. La non-reconnaissancee de ces syntagmes fournira certainement de faux résultats dans l'analyse de contenu basée sur la cooccurrence des syntagmes nominaux ;

3. Il faut encore étudier le comportement des verbes dans les syntagmes nominaux de manière à définir une syntaxe de réécriture de SN avec la présence de verbes, soit quand il apparaît seul dans le syntagme et aussi quand il apparaît accompagné des verbes auxiliaires. D'une certaine manière cette recherche est liée à la recherche des syntagmes nominaux avec des conjonctions de subordination et des pronoms relatifs puisque les verbes apparaissent souvent après ces dernières unités (la catégorie Q) ;

4. Une fois implémenté le modèle de système de recherche d'information proposé par cette thèse, il faut faire une étude d'évaluation de ce système. Pour cela il est nécessaire d'avoir un corpus suffisamment grand pour qu'on puisse évaluer le système proposé. Cette évaluation devra avoir comme but de connaître deux aspects du modèle proposé : 1) l'efficacité des syntagmes nominaux comme moyens d'accès à l'information ; 2) la satisfaction de l'usager en ce qui concerne la convivialité de l'interface de recherche d'information. Ainsi, il n'est pas suffisant d'avoir un bon corpus, mais il faut aussi constituer un ensemble d'usagers, organisé d'une telle manière qu'il puisse donner à connaître la performance de ce système pour les usagers novices et pour les usagers expérimentés. Il faut donc établir des critères bien définis autant au niveau du corpus qu'au niveau des usagers. Il ne faut pas oublier que ce modèle de système sera plus performant lorsqu'il travaille sur une base de données dans des domaines plus restreints. C'est-à-dire qu'il faut choisir un domaine d'information le plus spécifique possible, ce qui évitera les ambiguïtés ;

5. Une autre recherche qu'il faut développer, c'est l'évaluation ou plutôt la détermination des SN qui sont signifiants ou qui ont le statut d'un descripteur. Pour cela, il nous semble qu'il faut établir des critères qui puissent définir un descripteur et développer des outils de détermination des SN qui ont le statut de descripteur. Dans ce sens là, l'utilisation des outils statistiques (p. ex.: l'analyse factorielle et de correspondance)

peuvent aider dans cette procédure.

4 D'autres applications pour les Syntagmes Nominaux

Dans cette recherche nous n'avons travaillé que pour l'utilisation des SN comme moyens d'accès à l'information, comme descripteurs. Pourtant, l'utilisation des SN n'est pas restreinte à la recherche d'information : mais ils peuvent être utiles pour d'autres applications dans le champ de l'analyse de contenu et même dans les sciences de l'information et de la communication.

Aujourd'hui les outils pour l'analyse de contenu sont encore pauvres puisqu'ils ne travaillent que sur la base d'extraction et d'analyse de mots. Il existe des outils qui font des extractions d'unités plus complexes, mais ils font cela de manière encore primitive, c'est-à-dire de manière semi-automatisée, étant donné que ces outils demandent toujours des informations sur la composition de l'unité lorsqu'elle est complexe (p. ex.: nom + nom ; nom + de + nom ; etc.).

Dans la mesure où le modèle conçu dans cette recherche sera implémenté, le module de reconnaissance et d'extraction automatique des syntagmes nominaux peut remplacer ces outils utilisés dans l'analyse de contenu. Bien sûr, il faut encore ajouter quelques outils, puisque l'analyse de contenu ne se restreint pas à la simple extraction de mots, mais elle utilise aussi des outils de calculs statistiques pour déterminer les termes qui sont signifiants ou pertinents. Ainsi, il suffit d'intégrer ces outils statistiques au module de reconnaissance et d'extraction automatique des SN.

Aujourd'hui l'analyse de contenu est utilisée pour construire d'autres outils importants pour les sciences de l'information et de la communication et aussi pour la linguistique, parmi les applications couramment basées sur le mot, on distingue :

- dans le champ de la veille technologique ;
- dans la construction de vocabulaires préférentiels ;
- dans le champ de la terminologie.

Les thesaurus ont des procédures très lentes pour leur mise à jour, normalement on constitue des petits comités qui font l'analyse des termes candidats à être descripteurs dans le thésaurus. Le comité fait cela par le biais du recueil de termes candidats et de réunions périodiques où ces termes sont analysés. Or, la technologie aujourd'hui progresse de manière très rapide, et de même les termes de chaque domaine sont aussi renouvelés rapidement. Ainsi, les thesaurus ont une démarche de mise à jour très lente par rapport à la croissance ou au progrès du domaine de connaissance, et les thesaurus arrivent difficilement à suivre l'apparition de termes avec la même vitesse qu'elles apparaissent. Un des avantages d'utilisation des syntagmes nominaux est leur vitesse de mise à jour dans leur structure dans la base de données, puisque cela est fait au fur et à mesure que les documents sont mis dans la base de données. Et, si nous construisons

des outils statistiques analogue à ceux utilisés pour la détermination d'importance des mots, après chaque mise à jour de la base de données, nous pouvons avoir une relation de termes candidats au thesaurus avec leurs informations de pertinence ou d'importance. C'est-à-dire que ce petit comité peut se rassembler plus fréquemment et décider avec plus de sûreté l'inclusion ou non d'un terme ou syntagme nominal donné.

Ce sont quelques exemples d'utilisation des SN dans d'autres applications en dehors de la recherche d'information. Il nous semble que le champ d'utilisation des SN est encore ouvert, d'abord nous pouvons surtout supposer que les SN peuvent remplacer avantageusement la démarche basée sur le mot, souvent utilisée dans plusieurs types d'applications.

REFERENCES BIBLIOGRAPHIQUES

[ALBERICO et MICCO] ALBERICO, Ralph et MICCO Mary. *Expert Systems for Reference and Information Retrieval*. Westport : Meckler, 1990. 395 p. (Supplements to computers in libraries).

[ARNAULD & LANCELOT] ARNAULD, Antoine & LANCELOT, Claude. *Grammaire Générale et Raiosonnée de Port-Royal.* Généve : Slatkine Reprints, 1993.

[BACKUS] BACKUS, J. W. « The syntax and semantics of the proposed international algebraic language of the Zurich ACM-GAMM Conference ». In. : *Proceedings of International Conference on Information Processing.*UNESCO :1959, p. 125-132.

[BALPE] BALPE, Jean-Pierre. *Hyperdocuments, Hypertextes, Hypermedia*. Paris : Eyrolles, 1990. 200 p.

[BENVENISTE] BENVENISTE, Émile. *Problèmes de linguistique générale, 1.* Collection TEL. Éditions Gallimard, 1966. 356 p.

[BENVENISTE] BENVENISTE, Émile. *Problèmes de linguistique générale, 2.* Collection Tel. Paris : Editions Gallimard, 1974. 286 p.

[BERRENDONNER] BERRENDONNER, Alain. *Grammaire pour une analyseur :aspects morphologiques*. Les Cahiers du Criss. Grenoble : Centre de Recherche en Informatique appliquée aux Sciences Sociales. Université des Sciences Sociales de Grenoble, Novembre, 1990. 88 p.

[BINOT et al.] BINOT, J.-L., FALZON, P., PEREZ, R., PEROCHE, B., SHEEHY, N.,

ROUAULT J. et WILSON, M. « Architecture of a multimodal dialogue interface for knowledge-based systems ». In. : *Actes de la Conférence ESPRIT 90*. Novembre, 1990. p. 412-433.

[BINOT et al.] BINOT J.-L., DEBILLE, L., SEDLOCK, D., VANDECAPELLE, B. « Représentation Sémantique et Interprétation dans une Interface en Langage Naturel ». *Le Français Moderne*. Juin, 1991, t. LIX, n°. 1. p. 57-84.

[BOUCHÉ] BOUCHÉ, Richard. « Le Syntagme Nominal, une Nouvelle Approche des Bases de Données Textuelles ». *Meta*. 1989, vol. 34, n°. 3. p. 428-434.

[BOUCHÉ et al.] BOUCHÉ, Richard., LAINÉ, S. et 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.

[CALVINO] CALVINO, Italo. Leçons américaines :aide-mémoire pour le prochain millénaire. Gallimard : 1989, 197 p.

[CNPq] CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO (CNPq). INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA (IBICT). *Tesauro Ciência da Informação (Versão Preliminaire).* Brasília, 1989.

[CUNHA & CINTRA] CUNHA, C. et CINTRA, L. *Nova Gramática do Português Contemporâneo*. Lisboa : Edições João Sá da Costa, 1991. 734 p.

[DAMI et LALLICH-BOIDIN] DAMI, Samir et LALLICH-BOIDIN, Geneviève. « An Expert System for French Analysis within a Multi-mode Dialogue to be Connected ». In. : *RIAO 91 : Recherche d'Information Assistée par Ordinateur*. Barcelona, 1991. vol. 1. p. 431-451.

[DEWEZE] DEWEZE, A. *Informatique documentaire*. 4e éd. Paris : Masson, 1993. 292 p.

[DREYFUS] DREYFUS, Hubert L. *Intelligence Artificielle : mythes et limites.* Flamarion, 1984, 443 p.

[FLUHR] FLUHR, Christian. « Le traitement du langage naturel dans la recherche d'information documentaire». In.: *Cours INRIA - Interfaces Intelligentes dans l'Information Scientifique et Technique*. 18-22 Mai 1992. p. 103-128.

[GARY-PRIEUR] GARY-PRIEUR, Marie-Noëlle. « A propos du fonctionnement sémantique des noms propres et des noms abstraits. ». In. : Nelly Flaux, Michel GLATIGNY et Didier SAMAIN. *Les Noms Abstraits : histoire et théories.* Collection Sens et Structures. Paris : Presses universitaires du Septentrion, 1996. 406 p.

[GREVISSE] GREVISSE, Maurice. *Le Bon Usage*. 12ème édition. Editions DUCULOT, 1986. 1768 p.

[GRIES] GRIES, David. *Compiler Construction for Digital Computers.* New York : John Wiley & Sons, 1971. 493 p.

[GUIMIER-SORBETS] GUIMIER-SORBETS, Anne-Marie. « 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.

[HAASE] HAASE, A. *Syntaxe française du XVIIème siècle.* Paris : Delagrave, 1965.

448 p.

[HARTER] HARTER, Stephen P. *Online Information Retrieval : Concepts, Principles and Techniques*. Orlando : Academic Press. Inc., 1986. 259 p. (Library and Information Science).

[IHADJADENE] IHADJADENE, Madjid ; BOUCHÉ, Richard & KURAMOTO, Hélio. « Navegação nos vocabulários controlados ». À apparaître dans la revue *Revista Brasileira de Biblioteconomia.*

[JIN & FINE] JIN, Z. & FINE, S. « The Effect of Human Behavior on the Design of un Information Retrieval System Interface ». *Intl. Information & Library Revue.* Academic Press Ltd. : 1996, n. 28, p.249-260.

[KONG et SHOW] KONG, Hinny et SHOW, Guan Yeong. « Evaluation of parsing techniques for natural language processing ». In. : *Proceedings of the International Conference on Information for natural language processing*. Singapore, 1991. p. 422-432.

[KURAMOTO] KURAMOTO, Hélio. *Maquette d'un système de recherche d'information en utilisant des syntagmes nominaux.* Villeurbanne, 1995. Mémoire du DEA. École Nationale Supérieure des Sciences de l'Information et des Bibliothèques.

[LALLICH-BOIDIN] LALLICH-BOIDIN, Geneviève. *Analyse syntaxique automatique du français : Applications à l'indexation automatique.* Grenoble, 1986. Thèse de doctorat. Université des Sciences Sociales de Grenoble et Ecole Nationale Superiéure des Mines de Saint-Etienne.

[LANCASTER] LANCASTER, Frederic W. *Indexing and Abstracting in Theory and Practice.* London : Library Association Publishing Ltd., 1991. 328 p.

[LARDY] LARDY, Jean-Pierre. *Recherche d'Information dans Internet : outis et méthodes* . Paris : ADBS Editions. 3eme édition de mise à jour – Mai 1997, 118 p.

[LAROUK] LAROUK, Omar. *Extraction de connaissances à partir de documents textuels : traitement automatique de la coordination (connecteurs et ponctuation)*. Thèse de doctorat. Lyon : Université Claude Bernard – Lyon 1, 1993. 290 p.

[LARROCHE-BOUTET] LARROCHE-BOUTET, Valérie. Traitement linguuistique des anaphores possessives en indexation automatique : le cas des déterminants possessifs en français. Thèse de doctorat. Lyon : Université Lumière - Lyon 2, 1994.

[LE GUERN] 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*. Condenser, supplément I,Université Clermont Ferrand, 1982. 163-173 p.

[LE GUERN] LE GUERN, Michel. « Un analyseur morpho-syntaxique pour l'indexation automatique », *Le Français Moderne*. Juin, 1991, t. LIX, n°. 1, p. 22-35.

[LE GUERN] LE GUERN, Michel. « Traitement automatique et variation linguistique : la syntaxe des titres ». *Opérateurs et constructions syntaxiques : Evolutions des marques et des distributions du Xvème siècle.* Paris : Presses de l'Ecole Normale Supérieure, 1994. P. 75-81

[LE GUERN] 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.

[LEVY] LEVY, Piere. *L'Intelligence collective : Pour une anthropologie du cyberspace* . Paris : La Découverte, 1997. 246 p.

[MEKABOUCHE et BASSANO] MEKABOUCHE, A. et BASSANO, Jean-Claude. « Multi-experts Systems for Documentary research ». *RIAO 91 : Recherche d'Information Assistée par Ordinateur*. Barcelona, 1991. vol. 1, p. 394-413.

[MICROSOFT] MICROSOFT CO. *Microsoft Acces : Guide de l'utilisateur*. Ireland : 1994

[METZGER] METZGER, Jean-Paul. *Syntagmes Nominaux et Information Textuelle : reconnaissance automatique et représentation.* Thèse de Doctorat d'Etat en Sciences. Lyon : Université Claude Bernard – Lyon 1, 5 octobre 1988. 324 p.

[MINSKY] MINSKY, Marvin*. Semantic Information Processing*. Cambridge, Mass. : M.I.T. Press, 1969.

[MOHAMAD] MOHAMAD, Chawk*. La réécriture de D': les déterminants complexes du français : lexique et syntaxe*. Memoire de DEA, 1993.

[MOUNIN] MOUNIN, Georges. *Dictionnaire de la linguistique*. Paris : Quadrige / Presses Univesitaires de France, 1993. 340 p.

[POLITY] POLITY, Yolla. « Evaluation des modes de recherche en langage naturel ». *Documentaliste - Sciences de l'Information*. 1994, vol. 31, n°. 3. p. 136-142.

[POLLITT] POLLITT, Steven. « CANSEARCH : An expert systems approach to document retrieval ». *Information Processing and Management*. 1987, vol. 23, n°. 2. p. 119-138.

[SALTON] SALTON, Gerard. *Automatic Text Processing : The Transformation, Analysis and Retrieval of Information by Computer*. Massachusetts : Addison-Wesley Publishing Co., 1989. 530 p. (Computer Science).

[SALTON et MCGILL] SALTON, Gerard et MCGILL, Michael J. *Introduction to modern information retrieval*. New York : Mcgraw-Hill Book Company, 1983. 448 p. (Computer Science).

[SANDOVAL] SANDOVAL, Victor. *SGML : un outil pour la gestion électronique de documents*. Paris : Hermés, 1994. 174 p.

[SHOW et al.] SHOW Guan Yeong, KONG, Hinny et LIN, Kenneth Wente. « Intelligent user interface to SQL-based database system ». *Engineering Application Artificial Intelligence*. 1993, vol. 6, n°. 4. p. 307-316.

[SMEATON] 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, 21-23 Mars, 1989. p. 1-14.

[SMEATON] SMEATON Alan F. « Prospects for intelligent, languaged-based information retrieval ». *Online Review*. 1991, vol. 15, n°. 6. p. 373-382.

[SMEATON et SHERIDAN] SMEATON, Alan F. et SHERIDAN, Paraic. « Using Morpho-Syntaxique Language Analysis in Phrase Matching ». *RIAO 9 : Recherche d'Information Assistée par Ordinateur*. Barcelona, 1991. vol. 1, p. 414- 430.

[STRZALKOWSKI] STRZALKOWSKI, Tomek. « Natural language processing in large-scale text retrieval tasks ». *Text REtrieval Conference (TREC-1).* Gaithersburg,

1993. p. 173-187.

[TEYSSIER] TEYSSIER, Paul. *Manuel de Langue Portugaise : Portugal-Brésil*. Paris : Editions Klincksieck, 1984.

[WANG] WANG, Fangju. « Towards a natural language user interface : an approach of fuzzy query ». *International Journal of Geographical Information Systems.*1994, vol. 8, n°. 2. p. 143-162.

[WILMET] WILMET, Marc. « A la recherche du nom abstrait ». In. : Nelly FLAUX, Michel GLATIGNY et Didier SAMAIN. *Les Noms Abstraits : histoire et théorie.* Collection Sens et Structures. Paris : Presses Universitaires du Septentrion, 1996. 406 p.

[VAN HERWIJNEN] VAN HERWIJNEN, Eric. *SGML Pratique.*Paris : International Thomson Publishing France., 1995. 330 p.

[VAN HOE, R. ; POUPEYE, K. ; VANDIERENDONCK, A. ; et al] VAN HOE, R. ; POUPEYE, K. ; VANDIERENDONCK, A. ; et DE SOETE, G. « Some effects of menu characteristics and user personality on performance with menu-driven interfaces ». *Behaviour & Information Technology*. 1990, v. 9, n. 1, p. 17-29.

[VAN SLYPE] VAN SLYPE, George. *Les langages d'indexation : conception, construction et utilisation dans les systèmes documentaires*. Paris : Les éditions d'Organisation, 1987. 277 p.

[VIDALENC-SABOURIN] VIDALENC-SABOURIN, Isabelle. *Traitement automatique des anaphores en français : étude linguistique préalable.* Thèse de doctorat en Sciences de l'Information et Communication. Lyon : Université Lumière – Lyon 2, janvier 1989.

[VETTER] VETTER, Max. Modélisation des données : Approches globale et orientée objets. Paris : Dunod Informatique, 1992

BIBLIOGRAPHIE COMPLEMENTAIRE

[ADDISON et al.] ADDISON, E. R., WILSON, H. D. et FEDER, J. « The impact of plain English searching on end users ». In. : *Proceedings of the Fourteenth National Online Meeting.* New York, 1993. p. 5-9.

[ADDISON et NELSON] ADDISON, E. R. et NELSON, P. E. « Intelligent Hypertext ». In. : *Proceedings of the Thirteenth National Online Meeting*. New York, 1993. p. 27-30.

[AMARAL et SATOMURA] AMARAL, M. B. do et SATOMURA, Y. « Processing natrual languages at Chiba University Hospital ». *M Computing.* 1993, vol. 1, n°. 4. p. 6, 9-15.

[ARCHAMBAULT et BASSANO] ARCHAMBAULT, D. et BASSANO, Jean-Claude. « A neural network for supervised learning of natural language grammar ». *Proceedings of the Sixth International Conference on Tools with Artificial Intelligence. TAI 94*. New Orleans, 1994. p. 267-273.

[BASCH] BASCH, R. « Searching in plain English ». *Link-Up*. USA, 1994, vol. 11, n°. 2. p. 14-15.

[BASSANO] BASSANO, Jean-Claude. *DIALECT: un système expert pour la recherche documentaire*. Thèse de doctorat d'Etat. Paris : Université Paris 11, 1986.

[BELKIN et al.] BELKIN, (N. J.), COOL, C., CROFT, W. B., CALLAN, J. P. « The effect of multiple query representations on information retrieval system performance ». *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Pittsburgh, 1993. p. 339-347.

[BJORNER] BJORNER, S. « The . Where And . Why of FREESTYLE ». *Online*. 1994, vol. 18, n°. 2. p. 88-91.

[BLAIR] BLAIR, D. C. « Information retrieval and the philosophy of language : information retrieval ». *Computer Journal*. 1992, vol. 35, n°. 3. p. 200-207.

[BRAUNWARTH et al.] BRAUNWARTH, M., MEKABOUCHE, A. et BASSANO, Jean-Claude. « DIALECT-2 : an information retrieval system based on distributed artificial intelligence tools ». *Proceedings of the Sixth International Conference on Tools with Artificial Intelligence. TAI 94*. New Orleans, 1994. p. 800-803.

[BRAUNWARTH et al.] BRAUNWARTH, M., MEKABOUCHE, A. et BASSANO, Jean-Claude. « Using a dynamic blackboard model in a documentation retrieval system ». *Proceedings of Avignon '93. 13th. International Conference*. Avignon, 1993. vol. 3. p. 33-44.

[CARACENI et al.] CARACENI, R., GRAZIADIO, B., MUSSETTO, P., OBLIEGHT, A. et ZINNO, A. « Integrating data and text retrieval in a natural language system ». *Proceedings of Avignon '93. 13th International Conference*. Avignon, 1993. vol. 3. p. 55-64.

[CARENINI et al.] CARENINI, G., PIANESI, F., PONZI, M. et STOCK, O. « Natural language generation and hypertext access ». *Applied Artificial Intelligence*.1993, vol. 7, n°. 2. p. 135-164.

[CLEMSON] CLEMSON, P. A. « An approach to a networked document cataloguing. ». *Journal of Internet Cataloging.* v. 1, n. 2, 1997, p. 57-64.

[CLITHEROW et al.] CLITHEROW, Peter, RIECKEN, Doug et MULLER, Michael. « VISAR : A System for Inference and Navigation in Hypertext ». In. : *Hypertext'89. Proceedings. Special Issue - SIGCHI BULLETIN*. Pittsburgh, November 5-8, 1989. p. 293-304.

[COCH DI YACOVO et al.] COCH DI YACOVO, José, CRISPINO, Gustavo, CUKIERMAN, Diana, MORIZE, Geneviève et WONSEVER, Dina. « NAT-MULTILING, tools for multilingual interfaces with data bases ». *RIAO 91 : Recherche d'Information Assistée par Ordinateur*. Barcelona, 1991, vol. 1. p. 514-525.

[CORET et al.] CORET, Annie, MENON, Bruno, SCHIBLER, Daniele et TERRASSE, Christophe. « Un système d'indexation structurée à l'INIST. Bilan d'une étude préalable ». *Documentaliste - Sciences de l'Information*. 1994, vol. 31, n°. 3. p. 148-158.

[CROFT] CROFT, W. B. « The university of Massachusetts TIPSTER project ». *NIST special publication - TREC-1: Text Retrieval Conference.* Gaithersburg, 1992.p. 101-105.

[DAVIDSON] DAVIDSON, W. J. « SGML authoring tools for technical communication. ». *Technical Communication.* v. 40, n. 3, Août 93, p. 403-409.

[DE BERTRAND DE BEUVRON] DE BERTRAND DE BEUVRON, François. *Un système de progamation logique pour la création d'interfaces homme-machine en langue naturelle*. 1992. 293 p. Thèse de doctorat d'Etat, Université de Compiègne.

[DE BRITO] DE BRITO, Marcilio. Réalisation d'un analyseur morpho-syntaxique pour la reconnaissance du syntagme nominal : utilisation des grammaires affixes. Lyon,

1991. 221 p. Thèse de doctorat d'Etat, Université Claude Bernard, Lyon I.

[DE BRITO] DE BRITO, Marcilio. « Information System in natural languages: the search for an automatic indexing system ». *Ciência da Informação*. 1992, vol. 21, n°. 3. p. 223-232.

[DESCLES] DESCLES J. P. « Les interfaces homme-machine en langage naturel dans l'interrogation des bases de données relationelles ». *IDT. Information, documentation, transfert des connaissances*. Paris, 1994. p. 89-93.

[DESERT] DESERT, S. E. « WESTLAW is Natural v. Boolean searching: a performance study ». *Law Library Journal*. 1993, vol. 85, n°.4. p. 713-742.

[DESROCQUES et al.] DESROCQUES, Gilles, BASSANO, Jean-Claude et ARCHAMBAULT, Domenique. « An Associative Neural Expert System for Information Retrieval ». *RIAO 91 : Recherche d'Information Assistée par Ordinateur*. Barcelona, 1991. vol. 1. p. 546-566.

[DORWARD] DORWARD, A. « SGML in publishing : why use the standard ? ». *Electronic Library.* v. 13, n. 1, Fevrier 1995. P. 53-6.

[DRISCOLL et al.] DRISCOLL, J., LAUTENSCHLAGER, J. et MIMI, Zhao. « The QA system », *NIST special publication - TREC-1: Text Retrieval Conference*. Gaithersburg, 1992. p. 199-207.

[EVANS] EVANS R. « Beyond boolean: relevance ranking, natural language and the new search paradigm ». In. : *Proceedings of the Fifteenth National Online Meeting*. New York, 1994. p. 121-128.

[EVANS et LEFFERTS] EVANS, D.A. et LEFFERTS, R. G. « Design and evaluation of the CLARIT-TREC-2 system ». In. : *TREC-1: Text Retrieval Conference*. Gaithersburg, 1992. p.137-150.

[GALICY et al.] GALICY Jean Pierre, JOUIS, Christophe et GRAU, Brigitte. *SILNEBAD : Système d'interrogation en langage naturel d'un ensemble de bases de données*. 1989. 85 p. Rapport.

[GIROLLET et VICTORRI] GIROLLET, D. et VICTORRI, B. « The linguistic analyser of a smart gateway ». In. : *SEPLN - Sociedad Espanola para el Processamiento del Lenguaje Natural. VIII Congreso*. Granada, 1993. p. 29-39.

[GOLDFARB] GOLDFARB, Charles F. *The SGML Handbook.* New York : Yuri Rubinsky, Claredon Press – Oxford, 1990. 663 p.

[GUARDALBEN et LUCARELLA] GUARDALBEN, G. et LUCARELLA, D. « Information retrieval based on fuzzy reasoning », *Data & Knowledge Engineering*. 1993, vol. 10, n°. 1. p. 29-44.

[GRIFFITH] GRIFFITH, C. « Westlaw's WIN: not only natural, but new ». *Information Today*. 1992, vol. 9, n°. 9. p. 9-11.

[HAUSSER] HAUSSER, R. *Computation of Language*. Springer, Berlin, 1989.

[HERSH] HERSH, W. R. « SAPHIRE: a concept-based approach to automated indexing and retrieval in the biomedial domain ». *CURRENT RESEARCH 1992 National Library of Medicine Grant*.

[HESS] HESS, Michael. « An Incrementally Extensible Document Retrieval System Based on Linguistic and Logical Principes ». In. : *SIGIR'92*. Denmark, June, 1992. p.

190-197.

[JACSO] JACSO, P. « Don't kiss Boolean goodbye. It's AND not OR, let alone XOR ». *Information Today*. 1994, vol. 11, n°. 2. p. 22-24.

[KAHLE et al.] KAHLE, Brewster, MORRIS, Harry, DAVIS, Franklin, ERICKSON, Thomas, HART, Clare et PALMER, Robin. « Wide Area Information Servers : An Executive Information System for Unstructured Files ». *International High Speed Networks for Scientific and Technical Information (AGARD)*. 1993. p. 12/1-9.

[KUPIEC] KUPIEC, J. « MURAX: a robust linguistic approach for question answering using an on-line encyclopedia ».In. : *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Pittsburgh, 1993. p. 181-190.

[LANCEL] LANCEL, Jean Marie. « A pragmatic-based language understanding system (PLUS) ». *CEC-ESPRIT (Information Processing System).* November, 1990.

[LAROUK] LAROUK, Omar. « Linguistico-statistical approach and logic applied in documentary system ». *Proceedings of 8th SIGAPP Symposium on Applied Computing*. Indianapolis, 1993. p. 737-744.

[LARSON] LARSON, Ray R. « Classification clustering, probabilistic information retrieval, and the online catalog ». *Library Quartely*. 1991, vol. 61, n°. 2. p. 133-173.

[LASSALLE] LASSALLE, E. « Telmi: a reusable information retrieval system and its applications ». In. : *ASLIB Proceedings*. 1993, vol. 45, n°. 5. p. 144-148.

[LE LOARER] LE LOARER, Pierre. « OPAC: opaque or open, public, acessible and co-operative; some developments in natural language processing ». *Program*. 1993, vol. 27, n°. 3. p. 251-268.

[LEIGH] LEIGH, Sharon A. « The use of natural language processing in the development of topic specific databases ». In. : *Proceedings of the Twelveth National Online Meeting*. New York, 1991.p. 209-213.

[LIDDY et al.] LIDDY, E.D., WOOJIN, Paik et YU, E. S. « Text categorization for multiple users based on semantic features from a machine-readable dictionary ». *ACM Transactions on Information System*. 1994, vol. 12, n°. 3. p. 278-295.

[LIDDY] LIDDY, E. D. « An alternative representation for documents and queries ». In. : *Proceedings of the Fourteenth National Online Meeting*. New York, 1993. p. 279-284.

[LOSEE] LOSEE, R. M., Jr. « Term dependence: truncating the Bahadur Lazarsfeld expansion ». *Information Processing & Management*. 1994, vol. 30, n°. 2. p. 293-303.

[LUBKOV] LUBKOV, M. « Lexic: le langage naturellement ». *Archimag*. 1994, n°. 74. p. 33-35.

[MAREGA et PAZIENZA] MAREGA, Roberto et PAZIENZA, Maria Teresa. « CoDHIR : an information retrieval system based on semantic document representation ». *Journal of Information Science*. 1994, vol. 20, n°. 6. p. 399-412.

[MAULDIN] MAULDIN, Michael L. « Retrieval Performance in FERRET : A Conceptual Information Retrieval System ». *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development inf Information Retrieval*. Chicago, October 13-14, 1991. p. 347-355.

[MCCRAY et al.] MCCRAY, A. T., ARONSON, A. R., BROWNE, A. C., RINDFLESCH,

T. C. et SRINIVASIN, A. « UMLS knowledge for biomedial language processing ». *Bulletin of the Medical Library Association*. 1993, vol. 81, n°. 2. p. 184-194.

[MEKABOUCHE] MEKABOUCHE, A. *Un système multi-experts pour la recherche documentaire*. France, 1991. 263 p. Thèse de doctorat d'Etat, Univeersité d'Orleans.

[MENDIBOURE] MENDIBOURE, Catherine. *DOHQL: Une interface avancée pour l'interrogation de bases de données orientées-objet et documentaires*. Toulouse, 1994. 227 p. Thèse de doctorat d'Etat, Université de Toulouse 3.

[NELSON] NELSON, P.E. « The ConQuest system ». In. : *Text REtrieval Conference (TREC)*. Gaithersburg, 1993. p. 265-270.

[NELSON] NELSON, P.E. « Site report for the Retrieval Conference ». In. : *NIST special publication - TREC-1 : Text REtrieval Conference (TREC)*. Gaithersburg,1992. p. 287-296.

[NIE] NIE, Jian-Yun. « Towards a Probabilistic Modal Logic for semantic-based Information Retrieval ». In. : *15th Annual International SIGIR*. 1992. p. 140-151.

[OTT] OTT, N. « Aspects of the automatic generation of SQL statements in a natural language query interface ». *Information Systems*. 1992, vol. 17, n°. 2. p. 147-159.

[OWEI et HIGA] OWEI, V. et HIGA, K. « A paradigm for natural language explanation of database queries: a semantic data model approach ». *Journal of Database Management*. 1994, Winter, vol. 5, n°. 1. p. 18-30.

[PERNEL] PERNEL, Didier. Gestion des buts multiples de l'utilisateur dans un dialogue homme-machine de recherche d'informations. Paris, 1994. 328 p. Thèse de doctorat d'Etat, Université de Paris 11.

[PRITCHARD-SCHOCH] PRITCHARD-SCHOCH, Teresa. « WIN - Westlaw goes natural ». *Online*. 1993, Jan, vol. 17, n°. 1. p. 101-103.

[PRITCHARD-SCHOCH] PRITCHARD-SCHOCH, Teresa. « Natural Language Comes of Age ». *Online*. 1993, May, p. 33-43.

[PUGET] PUGET, Dominique. *Aspects semantiques dans les systèmes de recherche d'informations*. Toulouse, 1993. 183 p. Thèse de doctorat d'Etat, Université de Toulouse 3.

[PURDY] PURDY, W. C. « Alogic for natural language ». *Notre Dame Journal of Formal Logic*. 1991, vol. 32, n°. 3. p. 409-425.

[RABEN] RABEN, J. « SCHOLAR: implication for business ». In. : *Proceedings of the Fourteen National Online Meeting*. New York, 1993. p. 343-348.

[RADWAN et al.] RADWAN, Khaled, FOUSSIER, Frèderic et FLUHR, Christian. « Multilingual Access to Textual Databases ». *RIAO 91 : Recherche d'Information Assistée par Ordinateur*. Barcelona, 1991. vol. 1. p. 475- 489.

[RAMMAL] RAMMAL, Mahmoud. Une interface conceptuelle pour le traitement du langage naturel. Application au langage médical dans le système ADM. 1993. 202 p. Thèse de doctorat d'Etat, Universtié de Compiègne.

[ROWE] ROWE, Neil C. « Integrating depictions in natural-language captions for efficient access to picture data ». *Information Processing & Management*. 1994, vol. 30, n°. 3. p. 379-388.

[ROWE et GUGLIELMO] ROWE, Neil C. et GUGLIELMO, Eugene J. « Exploiting Captions in Retrieval of Multimedia Data ». *Information Processing & Management*. 1993, vol. 29, n°. 4. p. 453-461.

[ROWLEY] ROWLEY, Jennifer. « The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research ». *Journal of Information Science*. 1994, vol. 20, n°. 2. p. 108-119.

[SANDOVAL et al.] SANDOVAL, A. M., MORENO, C. O., GRISHMAN, R., MACLEOD, C. et STERLING, J. « PROTEUS: a multilingual system for information retrieval ». In. : *SEPLN - Sociedad Espanola para el Processamiento del Lenguaje Natural. VIII Congreso*. Granada, 1993. p. 47-56.

[SAMSTAG-SCHNOCK et MEADOW] SAMSTAG-SCHNOCK, Uwe, MEADOW, Charles T. « PBS : An Economical Natural Language Query Interpreter ». *Journal of the American Society for Information Science*. 1993, vol. 44, n°. 5. p. 265-272.

[SHULDBERG et al.] SHULDBERG, H. K., MACPHERSON, M., HUMPHREY, P. et CORLEY, J. « Distilling information from the EDS Template Filler System ». *Journal of the American Society for Information Science,* 1993. vol. 44, n°. 9. p. 493-507.

[SMEATON] SMEATON, Alan F. « Progress in the application of natural language processing to information retrieval tasks : information retrieval ». *Computer Journal*. 1992, vol. 35, n°. 3. p. 268-278.

[SUTCLIFFE] SUTCLIFFE, R. F. E. « PELICAN : a prototype information retrieval system using distributed propositional representation ». In. : *Proceedings of AI and Cognitive Science'91*. Cork, 1991. p. 147-163.

[TECNOPIR] TECNOPIR, C. « Online databases : natural language searching with WIN ». *Library Journal*. 1993, vol. 118, n°. 18. p. 54-56.

[THOMAZEAU] THOMAZEAU, Jacques. *Une interface multimodale pour l'interrogation d'une base d'objets complexes et documentaires*. Toulouse, 1993. 218 p. Thèse de docotorat d'Etat, Université de Toulouse 3.

[TRABELSI et al.] TRABELSI, Z., KOTANI, Y., TAKIGUCHI, N. et NISIMURA, H. « A database-domain hierarchy-based technique for handling unknown terms in natural language database query interfaces ». *IEICE Transactions on Information and Systems*. 1993, vol. E76-D, n°. 6. p. 668-679.

[TURNER et al.] TURNER, W. A., BUFFET, P. et LAVILLE, F. « LEXITRAN for an easier public access to patent database ». *World Patent Information*. 1991, vol. 13, n°. 2. p. 81-90.

[WIEDENHOLD] WIEDENHOLD, Gio. « Structural versus application knowledge for improved database interfaces ». In. : *Second Conference on Computer Interfaces and Intermediaries for Information Retrieval*. Alexandria, Defense Technical Info. Ctr., 1986. p.17-96.

[VEGA et OGONOWSKI] VEGA, J. et OGONOWSKI, A. « Natural-language access to the Dianeguide database », *SEPLN - Sociedad Espanola para el Processamiento del Lenguaje Natural. VIII Congreso*. Granada, 1993. p. 63-73.

[VICKERY et VICKERY] VICKERY, B. et VICKERY, A. « An application of language processing for a search interface ». *Journal of Documentation*. 1992, vol. 48, n°. 3. p. 255-275.

[VIEIRA] VIEIRA, Simone Bastos. La recuperación automática de información jurídica: metodología de análisis lógico-sintático para la lengua portuguesa. Madrid, 1994. 383 p. Thèse de doctorat, Universidad Complutense de Madrid.

[YOUNG et al.] YOUNG, Charlene W., EASTMAN, Caroline, M. et OAKMAN, Robert L. « An analysis of ill-formed input in natural language queries to document retrieval systems ». *Information Processing & Management*. 1991, vol. 27, n°. 6. p. 615-622.

[WALLACE] WALLACE, D.A. « Managing the present : metadata as archival description. ». *Archivaria.* v. 39 spring 95, p. 11-21.

[WARREN] WARREN, Karine. *Gestion de conflits dans une architecture multi-agents d'analyse automatique de textes.* Grenoble, 1998. Thèse de doctorat. Université Stendhal - Grenoble III - Equipe Cristal-Gresec.

[WILLIAMS] WILLIAMS, Martha. « The state of database today: 1996 ». *Gale directory of databases, volume 1: online databases, xvii-xxix.* New York: Gale Research Inc.

[WU] WU, Gilbert S. K. « SGML theory and practice ». *British Library*, 1989. 92 p.

[WU & ROBINSON] WU, G. & ROBINSON, B. « SGML support for secure document systems ». *British Library. Research and Development Department. BLRD Report.* 1994. 59 p.

[ZAMPOLLI et CALZOLARI] ZAMPOLLI, A. et CALZOLARI, N. « Linguistic tools for information retrieval. Documentary languages and databases ». *Advances in knowledge organization*. Rome, 1991. p. 174-200.

[ZWEIGENBAUM] 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.