Modèle d'analyse morpho-syntaxique adaptatif au web usages : ré-indexation sociale dans une norme syntagmatique

Sahbi Sidhom

Modèle d'analyse morpho-syntaxique adaptatif au web usages : ré-indexation sociale dans une norme syntagmatique

Sahbi SIDHOM (Université de Lorraine & lab. LORIA, France)
e-Mail : sahbi.sidhom@loria.fr

Résumé :

Le processus d'indexation consiste dans le fait de décrire des contenus dans une forme simple et manipulable pour les rendre exploitables et pour en assurer l'usage. Ce dernier est le plus habituel qu'est la recherche d'informations par le contenu. Ce dernier est décrit par une séquence structurée ou non de mots-clés (concepts ou descripteurs) ; cette séquence constitue l'index pour un document. L'usager quand il invoque le processus de recherche d'informations (RI), sa requête se formule en une séquence de mots empruntés ou assimilés au même vocabulaire que l'indexation, puis le système RI compare la requête et l'index des contenus pour proposer des documents qui coïncident en tout ou en partie au besoin informationnel exprimé.

Dans l'exposé de cette problématique, l'apport de l'usager, qui consulte des contenus tout en laissant des traces sur ses actions, permet de capitaliser des informations et des connaissances au profit du processus de réindexation. Ce dernier point ouvrira des réflexions de recherche autour du processus de gestion de contenus (voir le multimédia) par les usages et donc un réexamen sur l'organisation des connaissances entre les contenu, les usagers et les besoins.

Dans le contexte de la réindexation, il est nécessaire de repenser le système de management de la connaissance pour contenir les activités autour des contenus, usages et besoins. Pour se faire, la méthodologie proposée et appliquée se développe sur les aspects suivants : (i) Définition et filtrage de concepts sémantiques dans les contenus pour l'indexation, (ii) Organisation des connaissances dans le processus de recherche d'informations, (iii) Gestion de connaissances pour réponde à un besoin informationnel de l'usager, et (iv) Observations cognitives sur les contenus et implication des usages pour maîtriser la variabilité des implémentations formelles.

Mots-clés :

indexation automatique, observations cognitives, organisation des connaissances, filtrage sémantique, traitement automatique de la langue naturelle (TALN), réindexation, management de contenus.

1. Introduction

La diversité des applications réunies de nos jours sous les termes « industrie de la langue » et « industrie de la connaissance » recouvre plusieurs réflexions pour la recherche à l'ère de l'Internet ouvert aux usages et réseaux sociaux. Ce travail de recherche a consisté à baliser le terrain de ce que l'on convient d'appeler « traitement automatique de la langue naturelle » ou TAL à l'usage des ressources documentaires ouvertes sur le Web. Des expérimentations ont été construites, dans un premier temps, sur les documents audiovisuels de l'INA (www.institut-national-audiovisuel.fr) ont constitué en partie le corpus de travail et l'objet de notre étude, et dans un deuxième temps sur des corpus d'opinion sur le domaine de la santé et les nanosciences.

Pourtant, l'analyse linguistique automatique se trouve au confluent de plusieurs disciplines que sont la linguistique, la psycholinguistique, l'informatique et les mathématiques. La méthodologie dans ce travail inter-disciplinaire, nous autorise à prendre objectivement connaissance des problématiques épistémologiques dans chaque cadre d'objet d'étude. Nous nous servirons des concepts, des formalismes et des méthodes relativement concurrentes pour aborder la problématique de la ré-indexation des ressources et les usages.

Pour la qualité de l'étude, nous avons proposé une intégration de chaque objet d'étude et ses interactions (ou relations) avec les autres. En interaction linguistique-informatique, la présentation de la méthode nous permettra d'allier élégamment les concepts de chaque objet. Nous suggérons ensuite la construction de la structure morpho-syntaxique détaillant successivement la technique d'analyse par les syntagmes nominaux (ou SN) et leur mise opératoire dans un processus d'indexation automatique, tout en variant les ressources : du documentaire au web social.

Au début de notre étude, nous étions confrontés à un objet type qui est l'écrit comme résultat d'une production intellectuelle humaine. La forme de l'écrit est d'une grande variabilité, car elle est soumise à plusieurs facteurs extra-

linguistiques qui affectent aussi bien les conditions de production (connaissance, savoir et savoir-faire) que celles de leurs auteurs. La genèse même de l'écrit, à travers des études anthropologiques [GOODY, 94-98], a été soumise aux mêmes types de contraintes.

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

La recomposition de l'objet « ou la trace écrite » de l'auteur vers une forme stable offre un champ étendu pour l'interprétation. Le codage de la structure, qui une fois repérée et analysée, va nous permettre de tendre vers un de nos objectifs, à savoir l'analyse morpho-syntaxique automatique. Dans cette perspective de l'étude et le choix porté à un modèle linguistique calculable, il ne faut pas en effet perdre de vue que la qualité des résultats d'un analyseur placé dans un système de traitement automatique de la langue naturelle (TALN) puise ses performances, d'une part, de la qualité de sa conception (ou le formalisme d'implémentation), et d'autre part, de la qualité des recherches menées en syntaxe (ou autre modèle du langage théorique). Plus les concepts théoriques et pratiques seront en accord avec la nature de l'objet d'étude, et meilleures seront la qualité et l'efficacité des résultats de l'analyseur.

En particulier, l'étude menée sur le corpus de l'INA et l'étude linguistique fondée sur l'extraction des syntagmes nominaux et leurs propriétés, permettent d'observer des régularités structurelles et syntaxiques. Cette source de régularité était la base de la construction d'une grammaire formelle pour notre analyseur.

Explicitement, la démarche scientifique (*cf.* Fig.1.), que nous avons suivie pour étayer les hypothèses de notre travail et corréler nos choix théoriques avec nos conceptions pratiques, a consisté à :

(i) l'élaboration d'hypothèses sur les structures syntaxiques, qui se concrétise par l'étude linguistique sur le corpus : textes sur les analyses de contenu (INA) ;
(ii) la transcription des observations faites sur corpus en système stable de règles de réécriture grammaticale : formalisation ;

(iii) la matérialisation du système par l'implémentation de l'analyseur morpho-syntaxique ;

(iv) l'évaluation de l'analyseur par application directe sur corpus et sa comparaison aux observations retenues dans les hypothèses de l'étape (i) et la couverture de la grammaire formelle à l'étape (ii).

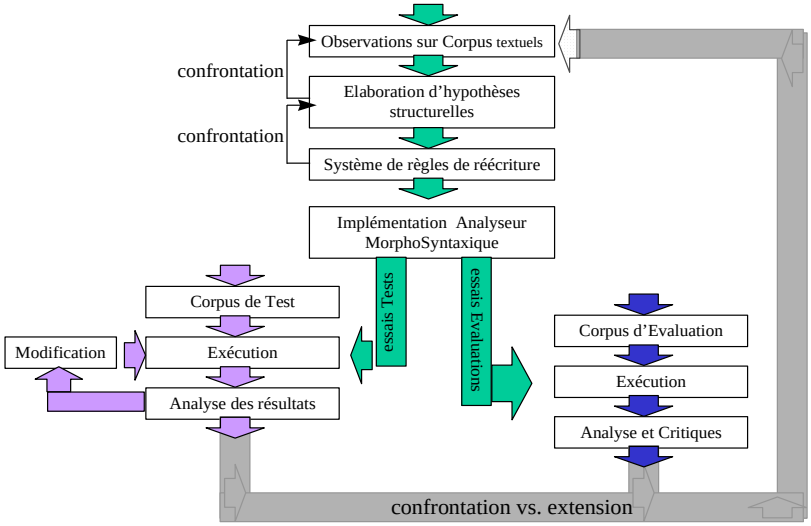

Fig.1. Démarche structuraliste sur l'objet d'étude.

Ainsi, l'objectif de notre recherche est d'aborder la question de l'organisation des connaissances, d'une part, par l'apport de l'indexation automatique et la recherche d'information sur les ressources informationnelles (banques de données documentaires, ressources ouvertes, etc.) et, d'autre part, par l'apport de la ré-indexation et les usages (web usages, enquêtes d'opinion, etc.) : principalement, la valeur ajoutée et les traces d'usage sur les contenus [PINON, 10-12].

Notre proposition est composée par la présentation d'un état de l'art sur l'indexation documentaire (en parag.2) qui argumente notre orientation vers la conception d'un modèle calculable (parag.3) fondé sur un *processus d'analyse du langage* qui utilise l'organisation des syntagmes nominaux comme descripteurs de l'information écrite. Dans l'enrichissement du modèle, le processus évolue vers l'organisation des connaissances (parag.4) fondée sur l'organisation naturelle des syntagmes nominaux et leurs propriétés sémantiques (relations

d'arbre, d'emboîtement et de classe) pour la recherche d'informations. Dans la logique de fonctionnement du modèle, la capitalisation des connaissances présente un outil orienté vers le management de la connaissance par l'implication de deux univers complémentaires (parag.5) l'indexation et la ré-indexation qui tendent vers l'usage pour la valorisation des contenus. Différents contextes d'étude (INA, santé et nanosciences) ont été présentés pour enrichir et valider nos hypothèses de travail

2. État de l'art sur l'indexation documentaire

Les bases de données documentaires ont l'ambition de mémoriser des informations sur les contenus (ou documents) en fonction de plusieurs critères (titre, auteurs, mots et descripteurs, thèmes, etc.) et dimensions (analytiques, descriptives, de contenus, etc.). Afin de répondre aux interrogations (ou besoins informationnels) des usagers, cette base leur fournit une sélection de documents pertinents [SIDHOM, 02].

En général, l'usager de la base ne connaît pas de références susceptibles de l'intéresser. Il essaye de formaliser sa demande lors de l'interrogation par le travail sur un thème générique ou spécifique selon des indicateurs référentiels ou thématiques. Par la suite, il constituera par la recherche d'informations un dossier rassemblant l'ensemble des concepts et mots s'y rapportant.

Ainsi, les systèmes d'information ou de recherche d'informations (documentaires) ont pour but de répondre à une telle demande d'informations en fournissant les documents adéquats et pertinents qu'ils retrouvent grâce à une indexation "judicieuse" [SALTON, 83-88].

L'opération d'indexation comme processus est particulièrement difficile dans la mesure où elle pose le problème de la représentation du sens dans un document. Dans ce cas précis, il faut souligner que les linguistes, les statisticiens, les analystes d'information et les informaticiens la traitent différemment.

2.1. Fondements théoriques des langages d'indexation

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

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

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

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

- **Langages d'indexation contrôlés :** ils se rapprochent des langages naturels. Un langage d'indexation « peu contrôlé » correspond aux descripteurs choisis librement pour représenter le contenu d'un document. A l'opposé, les langages d'indexation « plus contrôlés » se différencient nettement de la langue naturelle pour verser dans des langages d'indexation « post-coordonnés » ou « pré-coordonnés ».
- **Langages d'indexation coordonnés :** ils sont de deux types. Ceux d'indexation post-coordonnée qui sont la combinaison des descripteurs et qui se fait au moment de la recherche documentaire, au même titre que les thésaurus [BERRIAU, 03]. Quant aux langages d'indexation pré-coordonnée sont la combinaison des termes et qui est fixée au moment de l'indexation, tout comme les langages de classification et les langages en chaîne.

La problématique sur la classification des langages documentaires ne s'oppose pas en termes de répartition des unités descriptives dans les classes. Les unités descriptives en elles-mêmes sont plutôt le matériau qui sert aux regroupements sémantiques véhiculés par les contenus. Un même composant (ou unité descriptive) peut être utile à plusieurs groupements. Ainsi, pour mieux résoudre les problèmes liés à l'indexation et à la recherche d'informations, des méthodes ont été développées. Elles consistent à mémoriser les aspects les plus informatifs du contenu des documents [DEWEZE, 81-93] et la combinaison logique des descripteurs entre eux. Ce type de représentation va permettre, d'une part, d'améliorer l'ensemble des descripteurs caractéristiques du domaine traité pour l'indexation des documents ; et d'autre part, d'obtenir des réponses plus sensibles aux questions complexes par des combinaisons entre descripteurs.

La pluralité des solutions se résout d'un point de vue global, par la recherche de formes stables structurellement, inhérentes au domaine et fortes sémantiquement. Cela réduit, sans l'éliminer toutefois, l'incertitude quant à la

validité des autres langages de classification obtenus par des modèles du langage, probabilistes ou statistiques [AMAR, 97-08].

2.2- Orientation vers le processus d'indexation fondé sur des outils linguistiques

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

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

- **J. Maniez** dans ses travaux (de 1977 à 1993) soulève une question générale: quelles sont les différences et ressemblances nécessaires entre LN et LI ? l'élément commun est l'universalité de la fonction référentielle (la nature symbolique de signe). L'usager ne cherche pas de termes d'indexation pour eux-mêmes mais pour les documents dont ils représentent le sujet ;
- **J-C. Gardin** dans ses travaux (de 1973 à 1997) qualifie les LI par rapport aux LN de métalangage (langage ou système symbolique) pour exprimer le contenu d'un document rédigé en LN ;
- **R. Fugmann** dans ses travaux (de 1982 à 1993) considère les LI et les LN comme complémentaires. Les concepts individuels référant à un seul objet sont exprimés en LN par une seule expression lexicale. Par contre, les concepts généraux référant à une multitude d'objets sont souvent rendus en LN par plusieurs expressions lexicales (synonymes), voire par des expressions non lexicales (périphrases). Ce constat se porte dans le domaine scientifique où la terminologie est en retard sur les notions.

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

- Un élément de réponse a été apporté par la linguistique structuraliste (en références à N. Chomsky, C.J. Fillmore, B. Pottiers, J. Lyons, etc.) par la modélisation d'une structure profonde qui permet de rendre compte des structures de surfaces aussi diverses que sont les langues naturelles.

- Un autre élément de réponse a été apporté par la grammaire des cas de C.J. Fillmore qui permet de fournir une liste de catégories fondamentales utilisables pour la syntaxe des LI [MANIEZ, 77].

Dans une synthèse comparative, on pourra rapprocher la structure profonde d'un document (ou un texte donné) à la liste de ses index pour bien montrer que cette structure se situe déjà à un niveau de généralité plus élevé [SMEATON, 89-91a], [Chaudiron, 08].

3. Modèle d'indexation calculable

L'étude théorique d'un modèle linguistique sur lequel se fonde notre approche pour l'indexation doit posséder la propriété d'être calculable. Cette propriété conduit à la mise en équation des principales caractéristiques de la langue, et donc à une description algorithmique cohérente [GROSS, 96]. Le modèle linguistique doit avoir recours à un modèle formel et implémentable, résultant de l'interaction [JACQUEMIN, 00] entre les aspects linguistiques et algorithmiques : des représentations calculables aux descriptions [HABERT, 91-97], [JACQUEMIN, 06-03].

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

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

Étant donné que la construction des énoncés est importante pour notre étude, l'analyse morpho-syntaxique dont il sera question devrait permettre de repérer des expressions particulières : les syntagmes nominaux et ses propriétés inhérentes, et d'en faire usage dans le contexte de l'indexation automatique pour la recherche d'informations. Hors de ce contexte immédiat, nous pourrons projeter ce travail dans les processus de veille technologique, de réindexation par les usages et d'innovation par la découverte des signaux faibles.

3.1 Fonction référentielle : le syntagme nominal (SN) comme descripteur

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

dans un système d'information (documentaire), aussi appelé ***"descripteur"***. Selon W. M. Elhadi [MUSTAFA, 89-92], le **terme** ne peut être autre chose qu'un « terme préférentiel » que choisira le documentaliste parmi tant d'autres qui se trouvent être autant de candidats descripteurs. Le « descripteur » pourra être, dans une certaine perspective, le synonyme de « terme » :

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

En complément, il convient de distinguer les descripteurs aux « **mots du lexique** ». M. Le Guern dans ses travaux (1982 à 1994) a établi une comparaison entre le mot de la langue et le descripteur, et met en évidence la différence entre la « synonymie lexicale » et la « synonymie référentielle » :

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

Selon le même auteur, le syntagme nominal est L'unité minimale de discours qui a la possibilité de signifier un objet, ainsi :

> " {*MAISON*}*, le mot du lexique, ne signifie aucune maison que ce soit, alors qu'il suffit que le discours construise le syntagme* {*UNE MAISON*} *pour que soit désigné un objet concret. La fermeture du prédicat par le quantificateur* {*UNE*} *le transforme en terme.*" [LE GUERN, 89].

3.2 Fonction logique : aspects intensionnel et extensionnel en logique

Sur le plan logique, le cerveau humain a la possibilité de fonctionner selon deux systèmes différents : la logique intensionnelle et la logique extensionnelle. La logique intensionnelle a la particularité d'être une logique sans univers de référence, c'est le cas du fonctionnement du lexique d'une langue naturelle. Dès lors, le lexique devient un ensemble d'éléments qui ne sont pas en relation avec des objets (*cf.* Tab.1 et *Fig.2*).

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

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

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

Tab.1. Principes logiques : intensionnelle et extensionnelle.

3.3. Fonction computationnelle : reconnaissance du SN

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

Les mots de la langue ne signifient que des propriétés et jamais des entités ou objet du monde réel. Ils signifient des attributs et non des substances, tant qu'ils ne sont pas mis en oeuvre dans un univers de discours (*cf. Tab.1.*).

Cependant, l'analyse morpho-syntaxique a besoin de caractériser les unités linguistiques d'un texte en connaissant un certain nombre d'informations qui se rapportent à eux.

Ainsi, l'objectif d'un tel modèle linguistique est de permettre l'identification des syntagmes nominaux, tout en mettant en évidence la transition entre les mots du lexique (prédicats libres) et les syntagmes nominaux (prédicats liés) qui pointent sur des objets de la réalité extra-linguistique. Cette transition selon des principes computationnels (ou traitements linguistiques automatiques) s'effectue à travers la structure syntaxique qui reconnaît les SN. Selon R. Bouché [BOUCHÉ, 88-89], le modèle conçu a comme objectifs de :

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

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

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

  N' : niveau de transition entre l'intensionnel et l'extensionnel. Il s'agit de la prise en compte de l'univers du discours considéré. L'intervention de syntagmes prépositionnels, qui introduit des éléments dont on peut présupposer l'existence, définit une classe d'objets de la réalité extra-linguistique. N' reste un prédicat libre mais lié à une classe d'objet (celle de N").
  Exemples : (…)fenêtre de la maison[N'], *(…)maison de Jean*[N'], *etc.*

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

Pour l'aspect computationnel, le processus d'indexation et intrinsèquement celui de recherche d'informations sera évidemment centré sur la mise en oeuvre du mécanisme de référence à la réalité extra-linguistique : les objets du discours détectés. Une analyse identique à celle du document permet d'identifier dans la requête ou la question exprimée en langage naturel les syntagmes et les composants qui les lient.

3.4. Fonction cognition : préceptes d'identification d'une grammaire cognitive SN

Face à la prolifération des corpus d'étude et intrinsèquement la complexité dans la formalisation des grammaires d'analyse ou les difficultés liées à la capture des règles de réécriture, nous avons cherché à étudier la stabilité des descriptifs textuels. Tout particulièrement, ceux qui ont été développés dans les corpus INA (sources de INAthèque puis INAactualités) par les professionnels de l'audiovisuel. Cet examen est proposé afin d'établir par une analyse statistique les composantes grammaticales et syntaxiques de la phrase.

Lors de l'analyse, plusieurs situations se présentent où le repérage des syntagmes nominaux n'est pas toujours évident. Cela arrive parce qu'il y a des éléments anaphoriques, des ellipses, des syntagmes nominaux cachés, des syntagmes nominaux avec le déterminant zéro, etc.

Ainsi, il a fallu adopter quelques règles afin d'extraire les syntagmes nominaux de façon homogène pour obtenir des résultats statistiques cohérents dans un objectif précis : établir une grammaire de réécriture fondée sur les corpus. Tout en sachant que les corpus sont développés par des professionnels en texte libre et sans contraintes rédactionnelles.

Une manière de résoudre ces problèmes était de s'occuper seulement de l'extraction des syntagmes de surfaces « complets » sans traitement des cas anaphoriques, élliptiques, ou cachés. Seuls les SN avec déterminant zéro sont pris en compte, car nous supposons la facilité de remédier à ce type de problème lors de l'implémentation de l'analyseur morpho-syntaxique [SIDHOM, 02].

Le corpus de départ est constitué d'environ 100 notices bibliographiques sur des documents audiovisuels INA (sur des émissions radio et télévision) et qui a été étendu graduellement jusqu'à 300 notices. Chaque notice de l'INA contient au moins deux champs résumés (chapeau pour résumé synthétique et résumé pour les descriptions détaillées) produits par les professionnels [SIDHOM, 13].

En synthèse, l'analyse statistique sur le corpus de notices a révélé une stabilité grammaticale dans les descriptifs textuels (ou résumés de contenu). Cette révélation grammaticale cache en réalité une stabilité de rédaction des textes par les professionnels qui n'ont pas à priori de contraintes rédactionnelles, structurelles ou syntaxiques à respecter, si ce n'est que d'appliquer la grille d'analyse de contenu développée par type de source audiovisuelle.

La construction de la phrase (S) selon notre étude s'articule autour de trois structures fondamentales, à savoir : – une structure qui précède la phrase (une proposition introductive à S, notée PI), – le syntagme nominal sujet de S (sous forme d'un SN complexe), – le syntagme verbal de S (noté SV), et – la phrase relative (notée REL) en option. Chacune de ces structures est identifiée en ses éléments avec son organisation morpho-syntaxique composite :

S → [PI] + SN + [REL] + SV + [REL] ; [x] : structure optionnelle ;

Nous considérons que ce modèle de grammaire syntagmatique pour la phrase (S), d'une part, comme modèle « cognitif » pourra nous servir à la fois comme outil d'indexation ou un outil d'aide à la rédaction de textes et intrinsèquement l'orientation de son usage vers la réindexation sociale. D'autre part, même en variant le corpus de travail dans le contexte d'enquêtes d'opinion avec des questions ouvertes, les structures identifiées en ses éléments avec son organisation morpho-syntaxique se traduit par une sous-grammaire S' :

S' ⊂ S / S → [Vinf] + SN + [REL] ; [Vinf] : verbe-infinitif optionnel ;

4. Système d'organisation de la connaissance

4.1. Organisation morpho-syntaxique

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

Le contexte de coopération entre ces modules organisateurs est marqué par la conception d'outils informatiques nécessaires au traitement du langage naturel. Par conséquent, l'analyseur est destiné à opérer avec une grammaire. La

grammaire qui sera employée doit être celle d'un système orthographique où les régularités de l'écrit sont formalisées. La solution adoptée consiste à se donner un nombre très restreint de catégories syntaxiques, chacune ayant un comportement distributionnel bien défini.

Le prétraitement de nature morpho-syntaxique précède brièvement l'analyse morphologique dans le but de détecter, dans les séquences de formes, une propriété syntaxique quelconque. Par exemple l'occurrence de la forme {/ce/ + relatif} est de nature pronominale et non prédéterminative [DE BRITO, 91].

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

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

La grammaire s'exprime de façon normée au moyen de symboles et de règles. Les symboles terminaux sont des catégories morphologiques. Les règles peuvent faire intervenir, outre les catégories morphologiques, les variables associées à ces catégories pour compléter les conditions d'application de la règle. Cette grammaire a été proposée, pour l'essentiel, par A. Berrendonner pour l'analyse des SN. Elle a servi de support à plusieurs travaux de recherche effectués dans le cadre d'un groupe de chercheurs Lyonnais SYDO (pour SYstème DOcumentaire). Nous avons également retenu les améliorations apportées sur cette grammaire par les travaux de M. Le Guern [LE GUERN, 91, 94 ab] (fondements théoriques), J.-P. Metzger [METZGER, 85] (*réécriture du syntagme nominal*), M. De Brito [DE BRITO, 91] (*reconnaissance du syntagme nominal*), O. Larouk (*traitement de la coordination*) [LAROUK, 92], M. Chawk (*réécriture du déterminant complexe D'*) [CHAWK, 93], S. SIDHOM (*analyses de corpus pour la détermination des règles de réécriture cognitive du SN*) [SIDHOM, 98-99ab].

L'écriture de cette grammaire est inspirée de la notion X-barre, de N. Chomsky, pour représenter les structures syntaxiques organisées hiérarchiquement à partir de catégories principales : $X, \overline{X}, \overline{\overline{X}}$ (ou X, X', X''). Employer cette notion permet de générer les syntagmes principaux par l'emploi des deux règles :

$$(i): \quad \overline{\overline{X}} \rightarrow \quad spéc.\overline{X}\,\overline{X} \quad et \quad (ii): \quad \overline{X} \rightarrow \quad X$$

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

- $\mathbf{V_N}$: Vocabulaire Non-terminal de SN

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

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

Symbole	**Catégorie**
F-NOM	les noms
F-NOM-PRP	les noms propres
F-NOM-PRO	les noms pronoms
F-NAN	selon le contexte, nom ou adjectif
F-ADJ	les adjectifs
D	les prédéterminants
D-DEF	les prédéterminants définis
D-NUM	les prédéterminants numéraux , cardinaux et assimilées
D-IND	les autres prédéterminants
W-QUA	les adverbes de quantité
W-AAJ	les adverbes d'intensité (modificateurs d'adjectif)
P	les prépositions
P-DE	la préposition /de/
CI, LA	les mots /ci/ et /là/

- Quelques règles du syntagme nominal :

Description	N° règle	Règle
syntagmes nominaux :	1	N"→ D' + N + F-PRP
	2	N"→ D' + N'
	3	N"→ NOM-PRO
	4	N"→ NOM-PRP
expressions nominales :	5	N'→ N + SP + (SP)
	6	N'→ N + A'
	9	N'→ N
expressions prédéterminatives :	10	D'→ D-DEF + D-NUM
	11	D'→ P-DE + D-DEF
	13	D'→ W-QUA + P-DE
	14	D'→ D
centres adjectivaux :	15	A'→ W-AAJ + A
	16	A'→ A + EP
	18	A'→ A
centres nominaux :	19	N→ N + EP
	20	N→ N + A(QUA)
	21	N→ A(QUA) + N
nominaux :	22	N→ F-NOM
	23	N→ F-NAN
	24	A→ F-NAN,(QUA)
syntagme prépositionnel :	26	SP→ P + N"
expansion prépositionnelle :	27	EP→ P + N'

Cette grammaire de réécriture a été transcrite dans un formalisme ATN. Le formalisme ATN est formé de diverses classes de réseaux [WOODS, 80-86] qui correspondent aux classes des grammaires de la hiérarchie de N. Chomsky. Le réseau ATN (Augmented Transition Network) figure au sommet de la hiérarchie des réseaux. Il est obtenu à partir du réseau RTN (Recursive Transition Network), et d'un certain nombre d'ajouts (ou augmentations) permettant d'y intégrer l'équivalent de traitements réalisés par les grammaires transformationnelles. L'ATN est équivalent dans sa puissance de traitements symboliques à la Machine de Turing. A chaque réseau ATN est attaché un ensemble de registres précisant les *attributs* (exemple : genre et nombre pour un groupe syntaxique, etc.) et les *rôles* attachés à chaque structure engendrée (exemple, position sujet ou complément d'objet pour le groupe syntaxique).

4.2. Processus d'analyse

L'implémentation du processus d'indexation et intrinsèquement l'analyseur de la langue naturelle assurent la reconnaissance automatique des objets textuels dans un contenu en fonction des niveaux d'analyses ci-dessous :

- L'analyse morphologique : reconnaissance des lexèmes dans le lexique et la normalisation des mots fléchis dans les textes analysés.
- L'analyse syntaxique : construction de la représentation syntaxique pour chaque phrase ou segment de texte délimité. Les différents groupes syntaxiques de la phrase (ou segment) sont délimités ainsi que les relations entre ces groupes.
- L'interprétation sémantique : construction d'une représentation sémantique à partir de la représentation syntaxique précédente et la prise en compte du contexte : le syntagme nominal est l'élément clé dans cette représentation sémantique (structuration, représentation formelle, catégorisation conceptuelle).

Dans chaque chaîne analysée, l'analyseur construit une *série d'objets* comportant chacun la forme régularisée et son profil syntaxique, lexical et flexionnel. Ainsi, l'automate ATN lit successivement dans la série et selon le profil syntaxique de l'objet lu, passe ou non à un nouvel état.

Théoriquement l'automate ATN est caractérisé par les automatismes (ou machines) suivants :

1. Un automatisme qui traite la proposition principale et débute dès l'appel du type de la proposition déclenchée ;
2. Un automatisme sous-jacent et récursif à l'automate principal et se déclenche pour traiter les subordonnées de la proposition principale ;
3. Un automatisme ATN en cascade (ou CATN) permet de relier la sortie de certains automatismes comme entrée pour d'autres : ce principe d'automatisme permet la généralisation de la notion des ATN (machine type 1. ou 2.) [WOODS, 98].

Les mises en œuvre des automatismes ont été développées dans les constructions de l'analyseur (cf. détails dans [SIDHOM, 2002]).

4.3. Processus d'indexation

Les syntagmes nominaux ont une organisation naturelle. Dans un sens, ils ont un rapport d'emboîtement avec d'autres SN minimaux, ce qui permet de les classifier en des niveaux informationnels distincts. Et dans l'autre, ils ont un rapport d'arborescence, dans le cas où le syntagme nominal se présente avec une double rection. Cette dernière propriété permet d'ordonner des classes

d'informations : des structures d'arbre dans des classes d'information. Ces caractéristiques permettent de construire une architecture de connaissances et d'exploiter les primitives SN (emboîtement et arborescence) au moyen de la navigation.

Par la superposition des propriétés SN avec les centres de syntagmes (ou N), la navigation dans les structures s'intègre dans une architecture treillis de connaissances :

- Pour illustrer la caractéristique d'emboîtement, on présentera un exemple, d'un syntagme nominal de troisième niveau (*Fig. 2a*). On utilise le mot niveau pour indiquer l'ordre d'extraction des syntagmes nominaux. En effet, la grandeur du niveau est inversement proportionnelle à l'ordre d'extraction.

Exemple 1 : (*Fig. 2a*).

SN_{MAX} (la prise en charge de patients atteints de maladies chroniques) $\supseteq$ SN_1 (des patients atteints de maladies chroniques) $\supseteq$ SN_2 *(des maladies chroniques).*

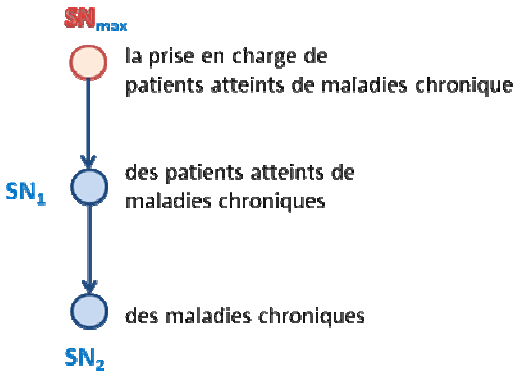

Figure 2a. Emboîtement de syntagmes nominaux.

- Pour illustrer la caractéristique d'arborescence, on présentera un exemple, d'un syntagme nominal avec une double rection (*Fig. 6b*). Nous présentons un exemple pour une meilleure visualisation de cette proposition.

Exemple 2 : (*Fig. 2b*).

SN_{max} *(le cadre du groupe de travail sur la prise en charge des patients atteints de maladies chroniques)* $\supseteq$ SN_{g1} *(le cadre du groupe de travail)* $\supseteq$ SN_{g2} *(le groupe de travail) AND*

SN_{max} *(le cadre du groupe de travail sur la prise en charge des patients atteints de maladies chroniques)* ⊇ SN_{d1} *(la prise en charge des patients atteints de maladies chroniques)* ⊇ SN_{d2} *(des patients atteints de maladies chroniques)* ⊇ SN_{d3} *(des maladies chroniques).*

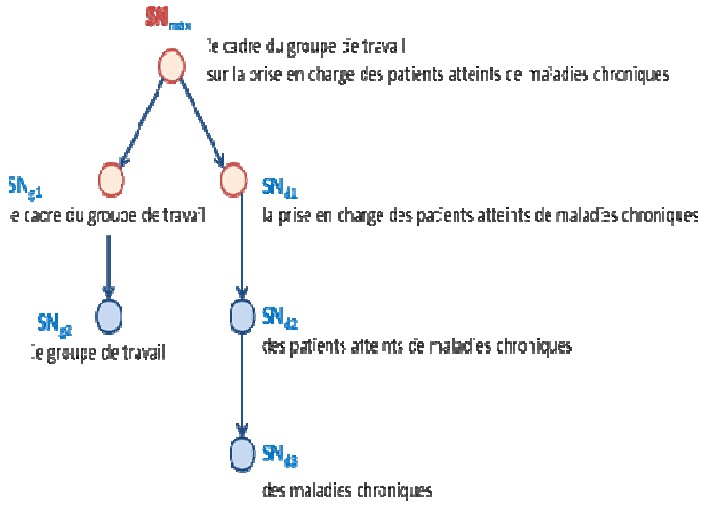

Fig. 2b. Arborescence de syntagmes nominaux.

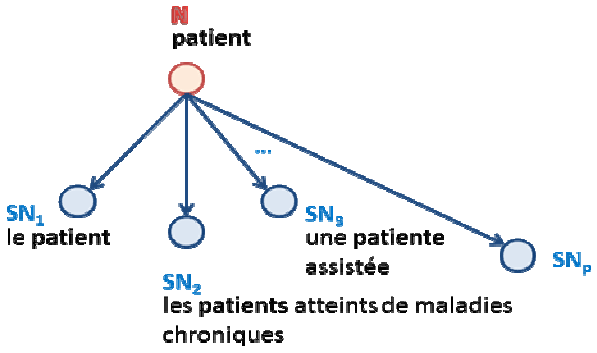

Fig. 2c. Classe de SN par le centre nominal N.

Etant donné que ce syntagme (*Fig. 2b*) contient une double rection (ou rupture sémantique), il est essentiel de repérer le niveau du syntagme nominal maximal par rapport aux syntagmes nominaux inférieurs de chaque rection.

- Le prédicat libre (N) dans le syntagme nominal est un élément appartenant à la logique intensionnelle. Ce prédicat N ne peut construire un objet de discours, mais comme trait d'une classe pour accéder aux éléments SN (*Fig. 2c.*). Ce prédicat est souvent représenté par un <*nom*> comme centre du syntagme nominal et contribuant à la description d'une classe d'objets (ou point d'accès).

Exemple 3 : (*Fig. 2c.*)

SN={*le patient, les patients atteints de maladies chroniques, une patiente assistée, etc.*} $\in N$=*patient.*

Ainsi, le rassemblement de tous les syntagmes nominaux et leurs propriétés dans une base de connaissances permettra de construire une structure treillis.

4.4. Organisation des connaissances SN

La différenciation des prédicats intensionnels (ou libres) aux prédicats extensionnels (ou saturés, les SN), permet de résoudre le problème majeur lié à l'extraction de l'information. La distinction des éléments, qui ont des propriétés prédicatives intensionnelles aux éléments qui ont des fonctions référentielles comme les syntagmes nominaux, permet de fournir une approche nouvelle de type référentielle (ou logique extensionnelle) dans le schéma de construction d'un système de recherche d'informations (*Fig. 3*).

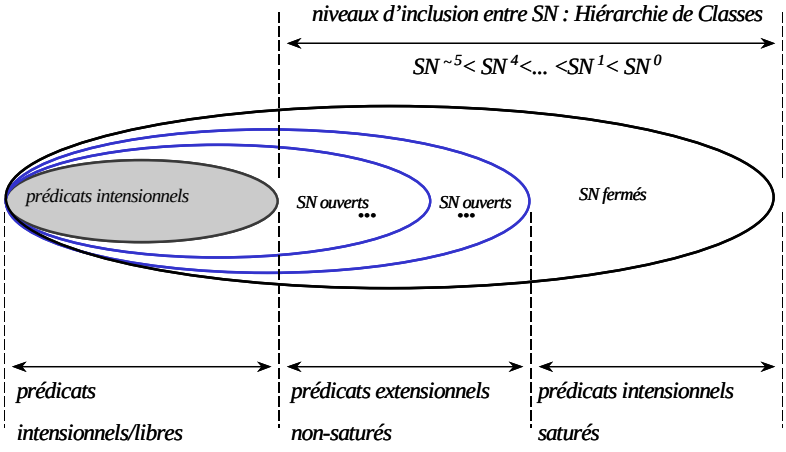

Fig. 3. Logique fonctionnelle d'un système de recherche d'informations.

Dans notre application, la base comportera des connaissances et des faits stockés :

- Prédicats intensionnels : nous avons vu qu'un SN peut se définir comme une suite de prédicats libres construits autour d'un nom N. Ce nom fait directement référence à un élément extra-linguistique. Le N employé comme centre du syntagme fera le lien à sa référence lors de son instanciation (ou saturation).
- Prédicats extensionnels non saturés : Le SN est presque toujours le thème en faisant référence à la correspondance entre les SN extraits d'un texte par l'analyseur et les descripteurs issus d'une indexation intellectuelle. Les SN non saturés correspondent aux SN emboîtés dans d'autres SN de niveau supérieur.
- Prédicats extensionnels saturés : Ils correspondent aux SN qui contiennent tous les autres SN de niveau inférieur. Un SN de ce type représente le thème complet dans le texte.

5. Management de connaissances : indexation et ré-indexation

5.1. Univers d'indexation

Une application multimédia est composée de données hétérogènes : textes, sons, graphiques, images fixes ou animations, vidéos. Dans leur organisation, en vue d'une veille documentaire ou informationnelle, les connaissances manipulables sont contenues dans les textes attachés à ces documents [BICHARD, 92] [MUSTAFA, 06-10], ou les parties composites du contenu.

La constitution d'un corpus audiovisuels associant des textes résumés (et annotations) pour notre étude, a pour but, dans une première étape de notre travail de recherche, d'opérer des analyses et de relever des régularités pour les règles syntaxiques. La nature même de ces résumés n'est pas construite ad-hoc, mais selon des principes fondés et une gestion d'analyse du contenu convenablement construits [CLAVEL, 93]. Cette procédure situe le sujet du document et ses différentes parties dans leur contexte.

Notre collaboration scientifique avec des spécialistes de l'INA (1998-2002) a révélé une expérience professionnelle et des acquis qui datent des années de création de l'ORTF. Depuis, la formulation des résumés de contenu sur des fonds documentaires hétérogènes est construite, dans certains organismes spécialisés, selon des critères et des méthodes formelles acquises par

l'expérience [BROWNE, 96] [MUCCHIELLI, 84]. Cela a permis dans notre étude de montrer une régularité et une constance des traitements réalisés par les professionnels de l'information [SIDHOM, 99AB]. Cette richesse documentaire (*cf.* exemple 1), une fois construite et mise à l'exploitation selon des traits attachés au contenu (capitalisation des sources d'information et de connaissances), pourra s'adapter aux diverses technologies d'exploitation et de diffusion [CHAMPENIER, 96], [PINON, 96], [MARET, 12A-12B-94].

Exemple 1 : Caractéristiques d'une notice (ex. INA) avec indexation de contenu.

Attributs	Exemples
Titre propre	1. Un lac venu de l'espace
Titre collection	Le monde des hélicoptères ;
Titre programme	Les cinq continents
Numéro	842.001
Numéro DL	DL T 19950101 FR2 022
	France 2
Producteurs	Producteur, Paris : France 3, 1994;Saint Ouen : Gédéon, 1994;Paris : ELF, 1994
Chapeau	Ce documentaire retrace les travaux menés par une équipe de chercheurs dans le nouveau Québec, afin d'expliquer la présence d'un lac qui se serait formé suite à la chute d'une météorite.
Résumé	La chute d'une météorite venue de l'espace a créé un lac dans la Toundra du Nouveau Québec. Celui-ci mesure 2,7 km de diamètre, 267 mètres de profondeur et son cratère s'étend sur 3 km.
Séquences	-DP Mirage en looping. -GP tête du pilote dans le cockpit. -Vieux coucou et Mirage faisant loopings côte à côte. -Auscultation par un médecin de JM Denuel. -Denuel se préparant au décollage, décollage. -Parachute s'ouvrant à l'atterrissage.
Résumé producteur	On ne réfléchit jamais assez aux miroirs, c'est bien connu. Une équipe de réalisation du C.N.R.S. s'est donc penchée, pour le compte de LA SEPT/ARTE, sur ces étranges surfaces à la fois aveugles et brillantes : on trouve de tout dans les miroirs et il y a des miroirs partout.
Notes de titre	Dépôt des cendres de Pierre et Marie Curie au Panthéon.

Notes	La version d'une heure de "Un lac venu de l'espace" a remportée le Prix de la meilleure vulgarisation scientifique au Festival International du film scientifique de Palaiseau en 1990.
Descripteurs	météorite; lac (Nouveau Québec); Québec; expédition (scientifique); chercheur
Descripteurs secondaires	sciences humaines; enseignement
Titre matériel	Un lac venu de l'espace : le cratère du Nouveau Québec
…	…

Le travail réalisé sur corpus INA et étendu à d'autres concerne l'identification des parties du discours construites autour du nom. Ces parties sont porteuses de références aux objets dans l'univers du discours. Elles sont celles identifiées aussi bien dans l'opération d'indexation que dans l'opération d'indexation de la requête de l'usager pour la soumettre au processus RI.

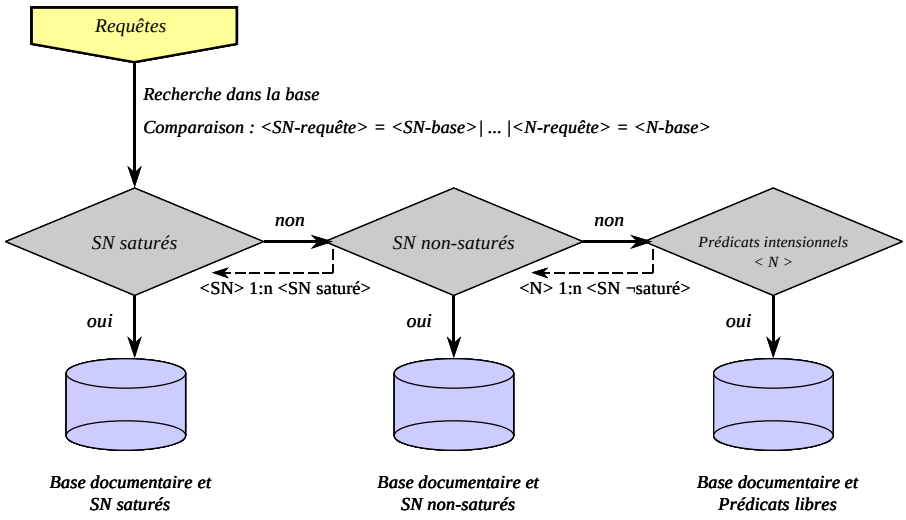

Fig. 4a. Schéma d'interrogation de la Base de connaissances.

Dans ce contexte, le schéma d'interrogation de la base des connaissances consiste à retrouver les SN de la requête qui sont présents dans la base. Bien entendu, les documents qui répondent le mieux à la requête sont ceux identifiés par des SN saturés, bien moins que par ceux identifiés par les SN non-saturés,

et encore moins par ceux identifiés par les prédicats intensionnels ou les nominaux (les classes de N) *Fig. 4a.*

Une information « *plus fine* » est la conséquence, d'une part, de la relation d'inclusion entre les syntagmes nominaux et, d'autre part, de la relation générique-spécifique entre le centre et son syntagme nominal. « *Plus riche* », elle est la conséquence du modèle linguistique (SYDO) adopté pour l'indexation documentaire fondée sur l'extraction des SN. Une description détaillée des structures morpho-syntaxiques du discours étant une bonne représentation de cette sémantique [SIDHOM, 11].

L'ensemble des associations reliant les références (SN) avec les traits (N) des classes d'objets, *c'est-à-dire* relations SN–SN et relations N–SN, décrit correctement l'organisation morpho-syntaxique des différentes parties du discours et permet d'établir un réseau cohérent d'informations structurées dans la base de connaissances.

Exemple 2 : Indexation des notices documentaires (INA)

Nous présentons dans le tableau suivant (*Tab. 2.*) un échantillon du fichier inverse de la base des notices documentaires. Les informations constituées permettent d'identifier les unités du discours de manière logique (les connaissances de type SN et leurs relations sémantiques, les classes N) en relation avec les unités documentaires (les notices DL_notice) avec le coefficient de la réussite de l'analyse du SN dans le contexte de sa production (coeff).

% filtrage	DL_notice	N	SN	SN^+	SN^-
83	137337.001	atelier	le atelier de construction aéronautique de Matra	_	Matra
92	87820.001	but	Le but de cette expédition	_	cette expédition
94	842.001	chercheurs	des chercheurs dans le nouveau Québec	une équipe de des chercheurs dans le nouveau Québec	le nouveau Québec

91	57399.001	compagnie	la compagnie de les dockers	le fonctionnement autogestionnaire de la compagnie de les dockers	les dockers
100	262714.022	développement	le développement de la intelligence artificielle à les robots industriels	_	la intelligence artificielle à les robots industriels
91	57399.001	dockers	les dockers	la compagnie de les dockers	_
94	842.001	équipe	une équipe de des chercheurs dans le nouveau Québec	les travaux menés par une équipe de des chercheurs dans le nouveau Québec	des chercheurs dans le nouveau Québec
92	87820.001	expédition	cette expédition	Le but de cette expédition	_
…	…	…	…	…	…

Tab.2. Echantillon du fichier inverse : connaissances SN.

5.2. Univers de ré-indexation

Pour un document donné, l'ultime élément informatif au sens où il renvoie à des éléments référentiels est de type SN ou N. Cependant, il est parfois difficile de faire coïncider les mêmes types d'objet dans l'univers de l'utilisateur par ses requêtes lors de la RI [WOODS, 98]. De même que ceux existant dans la base. Les rapprochements actuels des objets de l'utilisateur avec la base se font selon les niveaux logiques extensionnel (saturés ou non) puis intensionnel.

Le processus RI (*Fig.4b.*) consiste à rechercher les objets ayant des caractéristiques communes entre les requêtes et la base, *c'est-à-dire* appartenant à

une même classe ou sous-classe d'objets (les SN saturés et non saturés). Dans le cas échéant, ce processus consiste à retrouver des points d'accès communs aux classes d'objets, c'est-à-dire des prédicats intensionnels communs par les classes N.

Il est évident qu'un usager, qui a des idées précises sur le sujet de sa recherche d'informations, a une description qui ne se présente pas toujours sous la même forme [CROFT, 91] et cohérence de la base d'indexation à l'exception des objets uniques et universels qui appartiennent à tous les univers de discours. Cela est vrai en pratique pour les noms propres et de nombreux objets spécifiques. Il reste l'autre catégorie des objets non spécifiques.

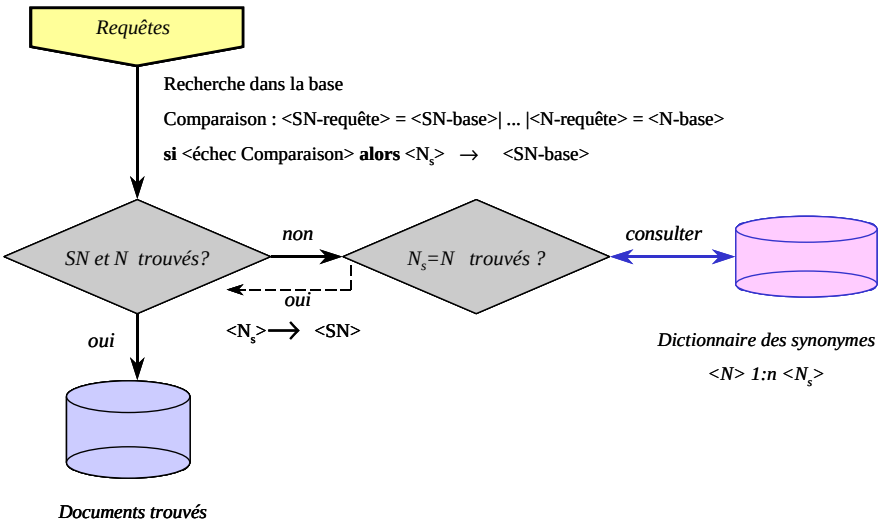

Fig. 4b. Schéma d'interrogation Requête et Base de connaissances.

Comment remédier à ce problème ?

La manipulation progressive des objets conduit à l'analyse des figures et des formes, tandis que la notion de l'objet demeure inchangée. Dès lors, le langage apparaît comme un prolongement de l'objet dans un espace qui lui accorde cette multiplicité (plusieurs descriptions du même objet). De l'avis de Saussure, « *dans la langue il n'y a que des différences.* » ; Dans ce cadre, l'unité linguistique n'est pas un signe, mais une valeur, et cette valeur est le résultat de relations complexes intervenant à plusieurs niveaux que celui d'une correspondance

simpliste entre signifiant et signifié à l'intérieur du mot et du morphème. Lorsque le signifiant et le signifié sont « pris séparément », seule leur combinaison « fondatrice du signe » est un fait positif : avec les règles de réécriture du modèle syntagmatique – analyse en constituants immédiats et règles engendrant l'indicateur syntagmatique de la structure profonde.

La première approche proposée (*Fig.7b.*) pour résoudre le problème d'échec de la recherche d'information consiste à expérimenter des recherches fondées sur les synonymes des prédicats intensionnels existant dans les requêtes.

Les synonymes Ns, qui sont des prédicats libres synonymes à ceux de la requête, permettront de retrouver dans la base les syntagmes ayant pour centre Ns et par conséquent les références aux documents de la base.

Exemple 1 : représentation étendue d'une requête

La requête peut se présenter (ou être formuler) comme le texte résumé dans une notice INA. Celui-ci a pour thème « Airbus industrie » (*réf.* Notice INA : DL T 19950107 M6 007) :

Titre	Airbus Industrie
REQ001	Pour les 20 ans d'AIRBUS INDUSTRIE, Frédéric BORSU retrace l'historique du consortium européen au moyen de documents d'archives. Du A 300B au A340, toutes les phases nous sont montrées pour comprendre les difficultés techniques et économiques qu'a pu rencontrer AIRBUS INDUSTRIE.
REQ002	AIRBUS INDUSTRIE parallèlement à l'Europe qui est en pleine mutation (chute du mur de Berlin, Tunnel sous la Manche...) prépare son avenir avec le A 321. Le A 300 B fit son vol inaugural, le 28 octobre 1972 avec Bernard ZIEGLER comme pilote d'essai. Mais toute la période de construction fut semé d'embauches tant économiques, politiques que techniques. La construction de la voilure fut construite à Chester en Angleterre, la section principale du fuselage à Hambourg en Allemagne, le cockpit et la section centrale à Saint-Nazaire et à Toulon par l'Aérospatial. Le plus dur fut d'acheminer les différentes parties à BLagnac pour l'assemblage du prototype. Les voies aériennes et routières fut empruntées non sans mal. Les tests draconiens du programme de certification sont passés avec réussite. Le succès commercial de l'Airbus A300B a permis au consortium européen d'élargir sa gamme d'appareils.

Suivant les mêmes techniques de recherche, de construction et d'assemblage, AIRBUS INDUSTRIE construira les modèles A310, A320, A330 et A340.

Exemple 2 : indexation automatique d'une requête

Le processus d'indexation de la requête (*Tab. 3.*) fait appel à notre analyseur morpho-syntaxique pour l'extraction des SN.

% filtrage	ref_req	N	SN	SN^+	SN^-
87	REQ002	A 321	le A 321	son avenir avec le A 321	_
33	REQ002	A 300	Le A 300	_	_
33	REQ002	Airbus	le Airbus A300B	Le succès commercial de le Airbus A300B	_
75	REQ001	ans	les 20 ans de AIRBUS	_	_
87	REQ002	avenir	son avenir avec le A 321	_	le A 321
100	REQ002	cockpit	le cockpit	_	_
94	REQ001	consortium	le consortium européen à le moyen de des documents de archives	le historique de le consortium européen à le moyen de des documents de archives	le moyen de des documents de archives
94	REQ001	historique	le historique de le consortium européen à le moyen de des documents de archives	_	le consortium européen à le moyen de des documents de archives

94	REQ001	moyen	le moyen de des documents de archives	le consortium européen à le moyen de des documents de archives	des documents de archives
100	REQ002	section	la section centrale à Saint-Nazaire	_	Saint-Nazaire
33	REQ002	succès	Le succès commercial de le Airbus A300B	_	le Airbus A300B
83	REQ002	voilure	la voilure	La construction de la voilure	_
…	…	…	…	…	…

Tab. 3. Echantillon du fichier inverse de la requête.

6. Conclusion

La contribution de ce travail s'inscrit au sein d'un domaine multidisciplinaire regroupant le traitement automatique du langage naturel, l'indexation associée à la recherche d'informations et l'organisation des connaissances à l'ère du numérique [COUZINET, 08-09], [Chaudiron, 10]. Sa particularité consiste en la mise à disposition d'outils pour le traitement automatique de l'information autour de l'écrit.

A ce titre, nous avons, dans un premier temps, clarifié l'Espace de Recherche dans lequel nous nous sommes situés. Nous avons tout d'abord posé les bases de notre discussion sur la connaissance écrite. Nous avons précisé les apports de cette connaissance qui ne se limite pas à l'écrit mais étendue à l'audiovisuel. L'idée de cette réflexion nous a amené à intégrer un nouveau composant concernant l'étude même de cet objet « connaissance » sur corpus. Le corpus concerné est un ensemble de notices documentaires (INA) puis étendu à des ressources sur le web [SIDHOM,11].

Le fruit de notre réflexion sur l'étude du corpus consistant à l'analyse des résumés a donné naissance à un modèle « cognitif » de rédaction. Cette révélation d'ordre grammaticale cache en réalité une stabilité de rédaction des

textes résumés que nous exploitons au profit des règles de réécriture de notre analyseur.

Le mécanisme d'analyse automatique s'est concrétisé par la conception d'un noyau d'indexation automatique qui a servi le processus RI. Ce processus permet, d'une part, le stockage des connaissances SN et ses constructions logico-sémantiques et, d'autre part, l'organisation des connaissances SN.

Cet aspect sur l'organisation des connaissances a été conduit dans le but de faire émerger les propriétés linguistiques et le traitement du langage dans une pratique expérimentale sur l'indexation automatique et la RI documentaire. Nous avons montré la nécessité de coordonner d'autres sources et stratégies dans l'exploration de ces propriétés. Il s'agit du mode de raisonnement et de la technique d'exploitation des objets du discours spécifiques à la gestion des connaissances. Ces deux derniers aspects intégrés dans le processus de l'organisation des connaissances offrent des scénarii pertinents pour la RI : la navigation dans le treillis des SN.

Dans une expérimentation en cours dans le domaine des nanosciences et nanotechnologies, avec l'implication des organismes de renoms en France (LNE, C'NANO, Club nanométrologie), elle s'avère intéressante en termes d'adaptabilité du formalisme, d'analyse et de recommandation. Des valorisations observées par la ré-indexation sociale au travers de nouveaux concepts sur des questions ouvertes qui ont fait l'objet d'un traitement automatique spécifique, à savoir : (i) « Quelles sont les raisons pour lesquels le répondant a adhéré au Club nanoMétrologie ? » et (ii) « Qu'est-ce qu'il attend spécifiquement d'une telle structure collaborative ? ». A l'issu des traitements et analyses du questionnaire qui a concerné une centaine de répondants. Des recommandations en matière d'aide à la décision ont pu être proposées pour le rapprochement des activités, des projets et des acteurs associant des compétences. Ces résultats soulignent la nécessité d'une activité de Community Management. L'intérêt de cette pratique renforce la proactivité des acteurs [LAMBERT et SIDHOM, 10] ainsi que leur cohésion pour l'émergence de nouveaux projets d'appels d'offre en nano. par le rapprochement en activités et des compétences [LAMBERT et SIDHOM, 11]. La méthodologie développée permet également le diagnostic de la structure interne du réseau Club nanoMétrologie. La détection de la nature hétérogène du réseau peut ainsi être mise à profit pour effectuer un ré-équilibrage autour du centre de gravité de la structure (ou la cohésion du réseau) : la cartographie des acteurs associés aux

thématiques du Club permet de mettre en évidence cette trilogie entre (A : acteurs, T : thématiques, R : ressources). Cf. Fig.5.

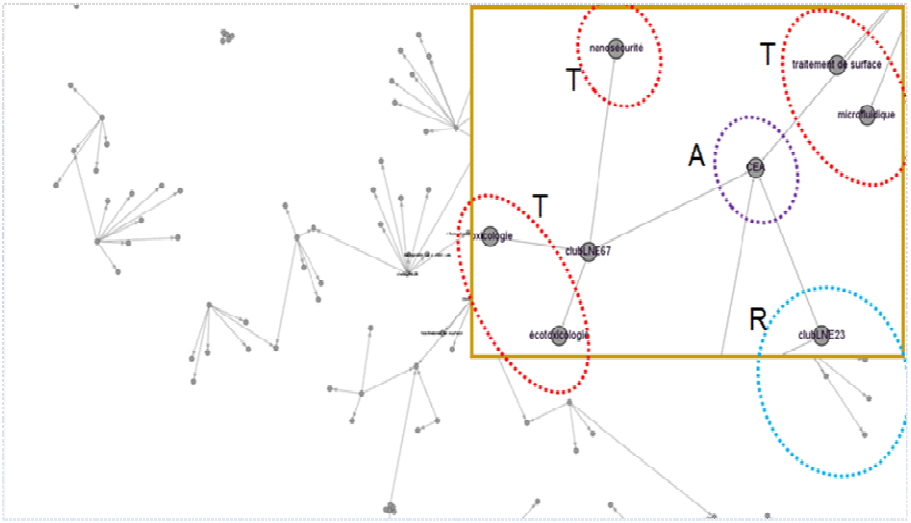

Fig. 5. Cartographie des acteurs associant thèmes et ressources du Club.

En perspectives, l'outil d'analyse en question, c'est-à-dire l'analyseur morpho-syntaxique, pourra servir comme un outil d'aide à la rédaction de documents ou de leur représentation textuelle et ainsi produire de manière systématique leur indexation.

Ainsi, on découvre l'organisation des différentes parties du discours au moyen d'une construction classificatoire. La classification en question est élaborée selon une organisation conceptuelle du SN. Celui-ci permet d'établir un réseau cohérent d'informations structurées et de construire de manière interactive un ordre opératoire à partir de la description de ses objets connaissances.

7. Bibliographie

[AMAR, 08] Amar, M., Camus-Vigué A., Evans, C., Gaudet, F. (2008) « Internet à la bibliothèque », In Sciences de l'information: problématiques émergentes, sous la dir. de Fabrice Papy, Paris: Hermès, p. 301-323.

[AMAR, 97] AMAR Muriel. Les fondements théoriques de l'indexation : une approche linguistique. Thèse de Doctorat en Science de l'Information et de la Communication : Université Lumière Lyon 2. 1997, 410p.

[BERRIAU, 03] BERRIAU, Nathalie et BELTRAME, Viviane .- Un outil d'aide à l'analyse documentaire : le thesaurus. In : Des nouvelles d'Ascodocpsy, numéro spécial, juin 2003, pp. 1-8.

[BICHARD, 92] J-P. Bichard. Les réseaux d'entreprise : accéder, partager, échanger. Paris : DUNOD, 1992.165p.

[BOUCHÉ, 88] Bouché R. .Valeur référentielle et langage d'indexation. Colloque Archives et temps réel, CEDRO-Université Lille III, Novembre 1988, Ed. ADBS Nord.

[BOUCHÉ, 89] Richard Bouché. Le syntagme Nominal - une nouvelle approche des bases de données textuelles. in META, 1989, vol. XXXIV, N°3, pp.428-434.

[BOUCHÉ, 90] BOUCHÉ R., LAINÉ S., METZGER J-P. Extraction des connaissances à partir d'une collection de documents. in : Tools of knowledge organization and the human interface, Congrès organisé par l'ISKO (International Society for Knowledge Organization), Darmstadt (D), 14-17 Août 1990.

[BROWNE, 96] Glenda Browne. Automatic indexing and abstracting. in Indexing in Electronic Age Conference, Robertson, NSW 20-21 April 1996, Australian Society of Indexers, 8p.

[CHAMPENIER, 96] Thiébaut Champenier, David Pautet. Mise à disposition à travers le réseau Internet de la littérature grise produite à l'INSA. Projet de PFE, 1996 , INSA-Lyon, 65p.

[CHAUDIRON, 08] Chaudiron S., Paroubek P., Hirshman L. (2008), Traitement automatique des langues, volume 47, n° 2, « Principes de l'évaluation en Traitement Automatique des Langues », ATALA, 2008.

[CHAUDIRON, 10] Chaudiron S., Ihadjadene M. (2010), « Electronic Information Access Devices : Crossed Approaches and New Boundaries », in Information Science, London, ISTE, p. 167-189

[CHAWK, 93] CHAWK Mohamad. La réécriture de D' - les déterminants complexes du français : lexique et syntaxe. Mémoire de DEA, enssib - Lyon, 1993, 108p.

[CLAVEL, 93] G. Clavel, F. Walther, J. Walther. Indexation automatique de fonds bibliothéconomiques. in ARBIDO-R, Vol.8 (1993) n°1, Suisse, p.14-19.

[COUZINET, 08] Couzinet V., Chaudiron S. (2008), Sciences de la société n°75, «Organisation des connaissances à l'ère numérique», Toulouse, Presses universitaires du Mirail.

[COUZINET, 09] COUZINET Viviane, 2009. Complexité et document : l'hybridation des médiations dans les zones en rupture, RECIIS, Electronic journal of communication information and innovation in Health, vol.3, n°3, p. 10-16.

[CROFT, 91] W. Bruce Croft, Howard R. Turtle, David D. Lewis. The Use of Phrases and Structured Queries. in Information Retrieval. SIGIR 1991: p.32-45.

[DACHLET, 90] Roland Dachlet. Etat de l'Art de la recherche en informatique documentaire : la représentation des documents et l'accès à l'information. Rapport de recherche de l'INRIA- Rocquencourt, Avril 1990, Projet : PSYCHO-ERGO - 32 p.

[DE BRITO, 91] De Brito M. .Réalisation d'un analyseur morpho-syntaxique pour la reconnaissance du syntagme nominal : Utilisation des grammaires Affixes. Thèse de Doctorat : Université Lyon 1, 1991, 220p.

[DEWEZE, 81] DEWEZE André. Réseaux sémantiques : essai de modélisation, application à l'indexation et à la recherche de l'information documentaire. Thèse de Doctorat : Informatique Documentaire : Université Claude Bernard Lyon 1, 1981.

[DEWEZE, 93] DEWEZE A. Informatique documentaire. 4e éd. Paris : Masson, 1993, 292 p.

[FUGMANN, 83] Fugmann, Robert. Analytico-synthetic foundation for large indexing & information retrieval systems. Bangalore : Sarada Ranganathan Endowment for Library Science, 1983 (OCoLC)566208602

[FUGMANN, 90] Fugmann, Robert. Tools for knowledge organization and the human interface (proceedings).International ISKO Conference (1st : 1990 : Darmstadt, Germany) Tools for knowledge organization and the human interface. Frankfurt/Main : INDEKS, 1990-1991 (OCoLC)551524447

[FUGMANN, 93] Fugmann, Robert. Subject analysis and indexing. Frankfurt am Main : Indeks Verlag, 1993 (OCoLC)606415067

[GARDIN, 73] J-C. Gardin. "Document analysis and linguistic theory", The Journal of Documentation, vol. 29, no 2, 1973, p. 137-168

[GARDIN, 74] J-C. Gardin. Les analyses de discours, Delachaux et Niestlé (Coll. Zethos), Neuchâtel, 1974

[GARDIN, 91] J-C. Gardin. La formalisation du discours savant : introduction à l'ouvrage Le calcul et la raison. Essais sur la formalisation du discours savant, Éditions EHESS, 1991.

[GARDIN, 97] J-C. Gardin. "Le questionnement logiciste et les conflits d'interprétation" : article de la revue Enquête, 5, 1997

[GOODY, 94] Jack Goody. Entre l'oralité et l'écriture. Traduit de l'anglais par Denise Paulme et révisé par Pascal Ferroli, Presses universitaires de France : PUF, 1994.

[GOODY, 98] Jack Goody. De l'oral à l'écrit. Propos recueillis par Nicolas Journet, in Sciences Humaines, Mai 1998, N°83, p.38-41.

[GROSS, 96] Maurice Gross, Max Silberztein. Outils de traitement linguistique, applications à l'analyse documentaire. Ecole d'été CNET (5, Trégastel 1995) – Traitement des langues naturelles, Editions CNET, p. I-1-24.

[HABERT, 91] Benoît Habert. Spécialiser des règles syntaxiques. 8eme Congrès AFCET – reconnaissances des formes et IA, Lyon novembre 1991, vol.2, Editions AFCET, p.873-878.

[HABERT, 97] B. Habert, M-L. Herviou-Picard, D. Bourigault, R. Quatrin, M. Roumens. Un outil et une méthode pour comparer deux extracteurs de groupes nominaux. Actes JST Francil 1997, Editions AUPELF-UREF, 1997 (Avignon) France, p.509-515.

[JACQUEMIN, 00] C. Jacquemin, P. Zweigenbaum. Traitement automatique des langues pour l'accès au contenu des documents. in Le document Multimédia en Sciences du traitement de l'Information, Ed. CEPADUES EDITIONS, Toulouse, Editors : J. Le Maitre alii, 2000, p.71-109.

[JACQUEMIN, 03] Jacquemin, C., and Bourigault, D. (2003). Term Extraction and Automatic Indexing. In R. Mitkov, editor, Handbook of Computational Linguistics, pages 599-615. Oxford University Press, Oxford.

[JACQUEMIN, 06] Virbel, J., Garcia-Debanc, C., Baccino, T., Carrio, L., Dominguez, C., Jacquemin, C., Luc, C., Mojahid, M., Péry-Woodley, M.-P., and Schmid, S., (2006). Approches cognitives de la spatialisation du langage. In, C. Thinus-Blanc and J. Bullier, editors, Cognitique, Agir dans l'espace, Editions de la Maison des Sciences de l'homme.

[LAMBERT et SIDHOM, 11] Lambert P., Sidhom S. (2011). Problématique de la veille informationnelle en contexte interculturel : étude de cas d'un processus d'identification d'experts vietnamiens". in Proceedings : ISKO-Maghreb'11 – Concept and Tools for Knowledge Management (KM). ESCE-University of la Manouba Edition. Hammamet (Tunisia) May. 2011.

[LAMBERT et SIDHOM, 10] Lambert P., Sidhom, S. (2010). Vers le Design d'information pour valoriser les résultats d'une veille sur les maladies chroniques. in Proceedings: Journée d'étude sur la "Mutualisation des ressources documentaires : Hétérogénéité des ressources et accessibilité dans un espace collaboratif." ELICO - Université Jean Moulin Lyon3, 05/11/2010 Lyon (France).

[LAROUK, 93] LAROUK Omar. Extraction de connaissances à partir de documents textuels : traitement automatique de la coordination (connecteurs et ponctuation). Thèse de doctorat: Université Claude Bernard Lyon 1, 1993. 290p.

[LE GUERN, 82] Le Guern M. Les descripteurs d'un système documentaire : essai de définition. In. : Bès, G.C., Fauchère, P.M., Lagueunière, F. Actes du Colloque Traitement automatique des langues naturelles et systèmes documentaires : Université Clermont Ferrand, 1982, p.163-173.

[LE GUERN, 89] Le Guern M. Sur les relations entre terminologie et lexique. in actes du colloque: les terminologies spécialisés - Approches quantitatives et logico-sémantique, et Meta Vol.34, No.3., sept. 89.

[LE GUERN, 91] Le Guern M. Un analyseur morpho-syntaxique pour l'indexation automatique. Le Français Moderne. Juin, 1991, tome LIX, n°. 1, p.22-35.

[LE GUERN, 94] Le Guern M. Parties du discours et catégories morphologiques en analyse automatique. Les Classes de Mots, Lyon : Presses Universitaires de Lyon, 1994, p. 207-215.

[MANIEZ, 77] Maniez J. Le rôle de la syntaxe dans les systèmes de recherche documentaire - Tome1: Aspects linguistiques, et - Tome 2: Etude critique de quelques SRD. IUT de Dijon- Département carrières de l'information, 1976-77, 184p. et 182p.

[MANIEZ, 87] Maniez J. Les langages documentaires et classificatoires : conception, construction et utilisation dans les systèmes documentaires. Paris : Editions d'Organisation. (Coll. Système d'information et de documentation), 1987.

[MANIEZ, 88] Maniez J. Relationship in thesauri : some critical remarks. International Classification, 1988, Vol. 15, n° 3, p.133-138.

[MANIEZ, 91] Maniez J., de Grolier E. A decade of research in classification. International Classification, 1991, Vol. 18, n° 2, p.73-77.

[MANIEZ, 93] Maniez J. L'évolution des langages documentaires. Documentaliste et Sciences de l'information, 1993, vol.30, n°4-5, p.254-259.

[MARET, 12A] Laurent Vercouter, Pierre Maret: Introducing Web Intelligence for communities. Web Intelligence and Agent Systems 10(1): 91-92 (2012)

[MARET, 12B] Johann Stan, Pierre Maret: Semantic metadata management in web 2.0. WIMS 2012: 6

[MARET, 94] P. Maret, J-M. Pinon, D. Martin. Capitalisation of consultants' experience in document drafting. Conference Proceedings RIAO 1994, Printed by CID Paris France, p.113-118.

[METZGER, 85] J-P. Metzger. Bases de données textuelles et analyse morphosyntaxique. in Textes des Communications IDT'85, Versailles 12-14 Juin 1985, Ed. adbs & anrt Paris, p.33-38

[METZGER, 88] J-P. Metzger. Syntagmes Nominaux et Information Textuelle : reconnaissance automatique et représentation. Thèse de Doctorat d'Etat en Sciences : Université Claude Bernard – Lyon 1, 1988, 324 p.

[MUCCHIELLI, 84] MUCCHIELLI, Roger. L'analyse de contenu des documents et des communications, connaissance du problème. Paris, Éditions ESF et Entreprise moderne d'édition, 1984. (ISBN 2-7101-0496-2)

[MUSTAFA, 06] Mustafa El Hadi, W. (2006) (dir) Terminologie et accès à

l'information, Hermès, coll. Traité des Sciences et des Techniques de l'information, Paris, 2006.

[MUSTAFA, 10] Mustafa El Hadi, W. Hudon, M. dir (2010), « Organisation des connaissances et web 2.0. » Numéro thématique des Cahiers du Numérique, LCN, VOL.6 N°3/2010, Paris Hermès-Lavoisier.

[MUSTAFA, 89] Mustafa Elhadi W. .La terminilogie arabe des télécommunications : Faits de variations. Thèse de Doctorat en Science du langage : Université Lyon2, 1989.

[MUSTAFA, 92] Mustafa Elhadi W. .La contribution de la terminologie à la conception théorique des langages documentaires et à l'indexation de documents. Meta, 1992,vol. XXXVII, n°3, p.465-473.

[PINON, 10] REIM Doumat, E. Egyed-Zsigmond, J.M. Pinon. User Trace-Based Recommendation System for a Digital Archive. Dans ICCBR 2010, Isabelle Bichindaritz, Stefania Montani ed. Alexandria, Italie. pp. 360-374. Lecture Notes in Computer Science 6176. Springer . ISBN 978-3-642-14273-4. 2010.

[PINON, 12] Modeling, Encoding and Querying Multi-Structured Documents. P-E. Portier, N Chatti, S. Calabretto, E. Egyed-Zsigmond, J.M. Pinon. Information Processing & Management, Elsevier. 2012.

[PINON, 96] Jean-Marie Pinon. Projet SEMUSDI : Serveur de documents Multimédia en Sciences de l'Ingénieur. Rapport de Présentation Technique : insa de Lyon, Juillet 1996, 15p.

[ROUAULT, 88] Rouault J. .Apport Contrainte et limites du langage dans le traitement automatique des langues. in Colloque Fribourg, Suisse, Mars 1988, 23p.

[SALTON, 83] G. Salton et M. J. McGill. Introduction to modern information retrieval. In McGraw-Hill, 1983.

[SALTON, 88] Salton G. and Buckley C. (1988) - Term Weighting Approaches, in Automatic Text Retrieval, Information Processing and Management, 24:5, 513-523.

[SIDHOM, 02] SIDHOM S. (2002). Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'information : de l'écrit vers la gestion des connaissances. Thèse de doctorat Université Claude Bernard - Lyon I (11/03/2002).

[SIDHOM, 11] Khemiri N., Sidhom S., Ghenima M., Ben Ghezala H.Natural ontology representation based on NP's properties and semantic relations. In Colloque International ISKO-Maghreb2011: Information Systems and Economic Intelligence 1 (2011) 315-321.

[SIDHOM, 13] Conjoncture des processus d'indexation et de gestion des connaissances : vers la réindexation par les usages. in Didactiques et métiers de l'humain et de la relation : nouveaux espaces et dispositifs en question. (direction de Frisch M.), ID Collection L'Harmattan. pp.85-125. Paris, 2013.

[SIDHOM, 98] SIDHOM S. Automatic indexing of multimedia documents based on the extraction of nominal phrases. Proceedings of 5th ISKO Conference, 25-29 august 1998 Lille, France, Ed. Ergon Verlag.

[SIDHOM, 99A] Sahbi Sidhom, Mohamed Hassoun, Richard Bouché. Cognitive grammar for indexing and writing. ISKO-España Conference Proceedings, 22-24 april 1999 Granada, p.11-16.

[SIDHOM, 99B] Sidhom S., Hassoun M., Bouché R., Colette Lustière, Daniel Gegez. Multimédia et exploitation textuelle pour un modèle d'indexation automatique. ISKO'99 Proceedings, Lyon France, 21-22 Octobre 1999.

[WOODS, 80] W. A. Woods. Cascaded ATN Grammars. in American Journal of Computational Linguistics, January-March 1980, vol.6, n°1.

[WOODS, 86] W. A. Woods. Transition Network Grammars for Natural Language Analysis. in : Natural Languge Processing, San Mateo : Morgan Kaufmann, 1986.

[WOODS, 97] W. A. Woods. Conceptual Indexing : a better way to organize knowledge. Technical Report SMLI TR-97-61 : SUN Micosystems, Lab. Mountain View Canada, April 1997.

[WOODS, 98] W. A. Woods, J. Ambroziak. Natural language technology in precision content retrieval. in Proceedings NLP+IA'98, August 1998, Moncton - New Brunswick CANADA