République algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université M'hamed BOUGARA de BOUMERDES
Faculté des sciences
Département Informatique

MEMOIRE

Présenté pour l'obtention du diplôme de Magister
Spécialité:
INFORMATIQUE
Option:
Spécification de logiciels et traitement de l'information

Par : **BOUCHAM Souhila**

Thème

Une approche basée Ontologies pour l'indexation automatique et la Recherche d'Information Multilingue (RIM)

Soutenu devant le jury:

Pr. M. Mezghiche	Professeur, Université de Boumerdes	Président
Dr. O. Nouali	Maître de recherche, CERIST	Examinateur
Pr. Alimzighi Zaia	Professeur, USTHB	Rapporteur
Mme Aliane Hassina	Chargée de recherche / CERIST	Invitée

Année universitaire 2008-2009

Résumé

Notre travail se situe dans le contexte de la *recherche d'information (RI)*, plus particulièrement la recherche d'information mulilingue (RIM).
L'objectif de ce projet est de proposer une solution pour la recherche d'information multilingue afin d'explorer l'apport des approches Web Sémantique en particulier l'utilisation des ontologies pour améliorer la description sémantique des documents et des requêtes.

Nous proposons dans ce travail une approche pour l'indexation et la recherche d'information pour un corpus trilingue : arabe, français et anglais. Le système proposé est fondé sur un formalisme de représentation de connaissances, plus précisément les graphes sémantiques qui supportent une ontologie de domaine. Les documents et les requêtes sont aussi représentés dans ce formalisme.
L'ontologie du domaine constitue le noyau du système et est utilisée aussi bien pour l'indexation que pour la recherche. Le système d'indexation utilise une méthode d'extraction qui est basée sur le calcul de segments répétés en utilisant des filtres linguistiques. Le système de recherche consiste en une comparaison de graphes pour trouver les documents qui répondent à la requête étendue de l'utilisateur.

Mots-clés : *recherche d'information multilingue, Indexation automatique, Ontologie, Extraction de connaissances, expansion de la requête, Graphes sémantiques.*

Abstract

This work deals with information retrieval (IR), more particulary multilingual information retrieval (MIR).
The aim our project is to propose a solution to MIR and explore the contribution semantic web approachs, particulary the the use of ontologies to improve the semantic description of documents and queries.

We propose an approach to indexing and retrieval in a trilingual corpus: arabic, french and english. The proposed system is founded on a knowledge representation formalism, namely semantic graphs which supports a domain ontology. Documents and queries are also represented in this formalism. The domain ontology constitutes the kernel of the system and is used both for indexing and retrieval.
The indexing system uses an extraction method based on repeated segments calculation. When identified, repeated segments are submitted to a filtering procedure using linguistic filters.
The retrieval system consists of a graph comparison to find relevant documents for the extended user query.

Keywords: multilingual information retrieval, automatic indexing, ontology, Extracting information, query expansion, semantic graphs.

ملخص

يدخل عملنا هذا في اطار بحث المعلومات وبالاخص بحث المعلومات المتعدد اللغات.
المشروع يهدف الى طرح حل لبحث المعلومات المتعدد اللغات من أجل استكشاف ما قدمه دنو الواب الالسني وبالاخص استعمال الانتلوجيات (علم الكائنات) لتحسين الوصف الالسني للمستندات (الوثائق) و الطلبات (الأسئلة).
نقترح في هذا العمل دنو للفهرسة و البحث المعلوماتي لقاعدة ثلاثية اللغات: عربية، فرنسية، انجليزية.
النظام المقترح يعتمد على تصميم و بنية تمثيل المعارف، وبالتحديد رسم بياني ألسني يستند على انتلوجيات المجال (الحقل).
الوثائق و الطلبات تمثل أيضا في هذا التصميم.
انتلوجيات المجال تكون أو تعتبر لب النظام كما تستعمل للفهرسة و البحث.
يستعمل نظام الفهرسة طريقة الاستخراج أو الاسترجاع التي تعتمد على حساب القطع المتكررة وهذا باستعمال مرشح لغوي.
نظام البحث يعتمد على مقارنة الرسوم البيانية لايجاد الوثائق المستجابة للطلب الموسع للمستعمل.

الكلمات الرئيسية: بحث المعلومات المتعدد اللغات، الفهرسة الأتوماتكية، الانتلوجيات، استخراج (استرجاع) المعلومات، الطلب الموسع، رسم بياني ألسني.

Remerciements

J'exprime mes grandes reconnaissances et mes vifs remerciements à la directrice de ma thèse le professeur Mme ALIMZIGHI Zaia, Maître de conférence, doyenne de la faculté d'électronique et d'informatique, USTHB.

Je remercie très vivement Mme ALIANE Hassina, chargée de la recherche / CERIST, qui m'a encadrée le travail durant le stage. Sa patience, ses encouragements et son écoute ont été d'un grand réconfort et d'une aide précieuse.

Je remercie très vivement le professeur MEZGHICHE, responsable de l'école doctorale en spécification de logiciels et traitement de l'information, pour les efforts qu'il a bien voulu consacrer à ses étudiants.

Je remercie vivement les membres de jury pour avoir accepté de juger ce modeste travail.

Enfin, je remercie tous ceux qui de près ou de loin ont bien voulu m'encourager pour que ce travail puisse être achevé.

A mes parents qui ont toujours encouragé ma curiosité intellectuelle,
A mon mari et mon bébé Ranim Riaâl,
A toute ma famille et tous mes amies.

Table des matières

Partie II

Introduction Générale

L'émergence de l'Internet a profondément transformé les moyens de communication, notamment en facilitant les échanges de documents entre les pays. Dès lors, les collections de documents se sont enrichies par des documents écrits dans différentes langues. Les systèmes de recherche d'information (SRI) ont dû s'adapter à cette révolution technique pour devenir des systèmes capables de gérer des collections multilingues de documents.
La recherche devenant donc multilingue : il faut retrouver tous les documents relatifs à une requête donnée quelque soit leur langue.
Par ailleurs, un SRI multilingue doit aussi faire face au problème de la représentation du contenu des documents ainsi qu'au problème de l'évolution de la pertinence. Cette évaluation est plus difficile que dans un SRI monolingue, en effet il est difficile de construire une fonction de correspondance avec différents langages pour les documents et la requête [Bao-Quoc, 04][Aliane et al, 06]

Différentes méthodes ont été proposées dans la littérature, L'objectif de ce travail est d'explorer l'apport des approches Web Sémantique pour cette problématique en particulier l'utilisation des ontologies pour améliorer la description sémantique des documents et des requêtes.
De là, la conception et le développement de notre système procèdent en deux étapes :

- La première construit en collaboration avec un expert humain une ontologie fondée sur le formalisme des graphes sémantiques et qui explicite les concepts du domaine et leurs relations.
- Dans la seconde étape, cette ontologie est considérée comme un bootstrap (une liste initiale de concepts du domaine) qui initialise le système de connaissance, alors, le processus d'indexation est basé sur une méthode linguistique de surface, plus précisément un algorithme d'extraction de segments répétés.

Pourquoi le multilingue : l'internautes en 2001, 45 % Anglais, 9,8 % Chinois et japonais et 29,8 % dans l'une des langues européennes, mais en 2005 29 % anglais.

Si l'utilisateur ne dispose que de SRI monolingues et qu'il veut récupérer les documents écrits dans d'autres langues, il doit traduire sa requête. Il doit ainsi soumettre autant de requêtes que de langues qu'il souhaite prendre en compte. Cette manipulation est lourde et n'est pas à la portée des usagers qui ne connaissent que leur langue maternelle.
Pourtant ces derniers peuvent s'intéresser à des documents écrits dans d'autres langues. Il suffit de penser au contexte de la veille stratégique où l'utilisateur cherchera à s'informer de façon détaillée sur un sujet particulier sans se limiter aux documents rédigés dans sa langue.
Une fois les documents retrouvés, pour accéder à leur contenu, l'utilisateur pourra utiliser un système de traduction automatique ou demander l'aide de traducteurs humains.
Il existe un besoin pour un système qui peut tenir compte non seulement du mot clé entré pour la recherche, mais également de la signification ou *concept* que l'utilisateur cherche.

Le travail que nous présentons ici essaie de prendre en considération ces aspects en se basant sur les ontologies pour représenter aussi bien l'information (souvent des documents textuels) que le besoin en information de l'utilisateur (requête). Les deux principales questions auxquelles nous essayons de répondre dans ce mémoire sont alors :

- Comment les approches web sémantique notamment les ontologies peuvent apporter un plus au domaine de la RIM ? .
- Comment peuvent elles être intégrées dans les processus de représentation et de recherche de l'information en RIM ?

Donc, notre étude se situe à la jonction des deux problématiques : la recherche multilingue et l'indexation à base de connaissances du web sémantique.

Organisation du Mémoire :

Le mémoire est organisé en deux parties principales : la première partie comporte trois chapitres sur l'état de l'art en lien avec le cadre de notre mémoire: les notions et concepts de base de la RIM, les ontologies et leur utilisation en RI.

La deuxième partie regroupe le détail de notre contribution.

Chapitre 1 : De la Recherche d'Information (RI) à la Recherche d'Information Multilingue (RIM)

1. La recherche d'information :

1.1. Introduction:

La recherche d'information est un domaine de la technologie d'information qui consiste à chercher sur une grande masse d'informations les documents qui satisfont les besoins d'utilisateur. Hors la quantité d'information stockée au format électronique ne cessant de croître, il devient de plus en plus difficile de retrouver un ensemble d'information contenu dans un document, au sein d'une base de documents, appelée corpus. De plus, l'information disséminée dans un document n'est pas structurée et donc difficilement accessible voire identifiable. Outre le problème d'identifier l'information contenue dans un document, la recherche d'information doit également permettre à l'utilisateur de formuler sa demande, son besoin d'information, le plus exactement possible, sous la forme d'une requête normalement en langage naturel.
Ce chapitre est consacré à la recherche d'information dans les documents, connue aussi sous le nom recherche documentaire. Nous intéressons uniquement à la partie textuelle des documents.

1.2. Définition d'un Système de Recherche d'Information :

Il existe plusieurs définitions d'un SRI, qui sont plus au moins proches :
Tomek Strzalkowski définit un SRI comme suit [Strzalkowski , 93] :
La tâche typique de la recherche d'information est de sélectionner des documents dans une base de données, en réponse à une requête de l'usager, et leur rangement par ordre de pertinence.
Tandis que Alan Smeaton donne la définition suivante [Smeaton, 89]:
Le but d'un système de recherche d'information est de retrouver des documents en réponse à une requête des usagers, de manière à ce que les contenus des documents soient pertinents au besoin initial d'information de l'usager.
Salton et Mcgill donnent une définition d'un SRI plus précise et complète [SM, 83]:
Un SRI traite de la représentation, du stockage des informations, de l'organisation de ces informations (processus d'indexation) et de l'accès aux éléments de l'information.

On définit un SRI comme étant un système permettant de retrouver les documents pertinents à une requête d'utilisateur écrite dans un langage libre, à partir d'une base de documents volumineuse (Figure I.1).
Dans cette définition, il y a trois notions clés : documents, requête, pertinence.

- *Documents* : un document peut être un texte, un morceau de texte, une page WEB, une image, une bande vidéo, etc. On appelle document toute unité qui peut constituer une réponse à une requête d'utilisateur. Dans cette thèse, nous traitons seulement des documents textuels.
- *Requête* : une requête exprime le besoin d'information d'un utilisateur.

- *Pertinence* : le but de la RI est de trouver seulement les documents pertinents. La notion de pertinence est très complexe. De façon générale, dans un document pertinent, l'utilisateur doit pouvoir trouver les informations dont il a besoin. C'est sur cette notion de pertinence que le système doit juger si un document doit être retourné à l'utilisateur.

1.3. Architecture générale des SRI :

Le processus de recherche d'informations pertinentes que le SRI est sensé restituer à un utilisateur, consiste en la mise en correspondance des représentations des informations contenues dans un fond documentaire des besoins de cet utilisateur exprimés par une requête. Cette notion de pertinence peut être appréhendée à deux niveaux :

- *Niveau utilisateur* : la pertinence correspond à la satisfaction de l'utilisateur de l'ensemble des documents restitués par le SRI.
- *Niveau système* : le système mesure un degré de pertinence, une valeur de similitude entre un document et une requête.

Le but de tout SRI est de rapprocher la pertinence système de la pertinence utilisateur. Pour effectuer de façon efficace cette fonction, le SRI doit réaliser l'indexation des documents, la formulation de la question, la comparaison question-documents et enfin la reformulation de la requête (processus non toujours présent mais important).
Nous pouvons représenter schématiquement un SRI, comme illustré par la figure 1.1, par ce qui est appelé communément le processus en U de recherche d'information.

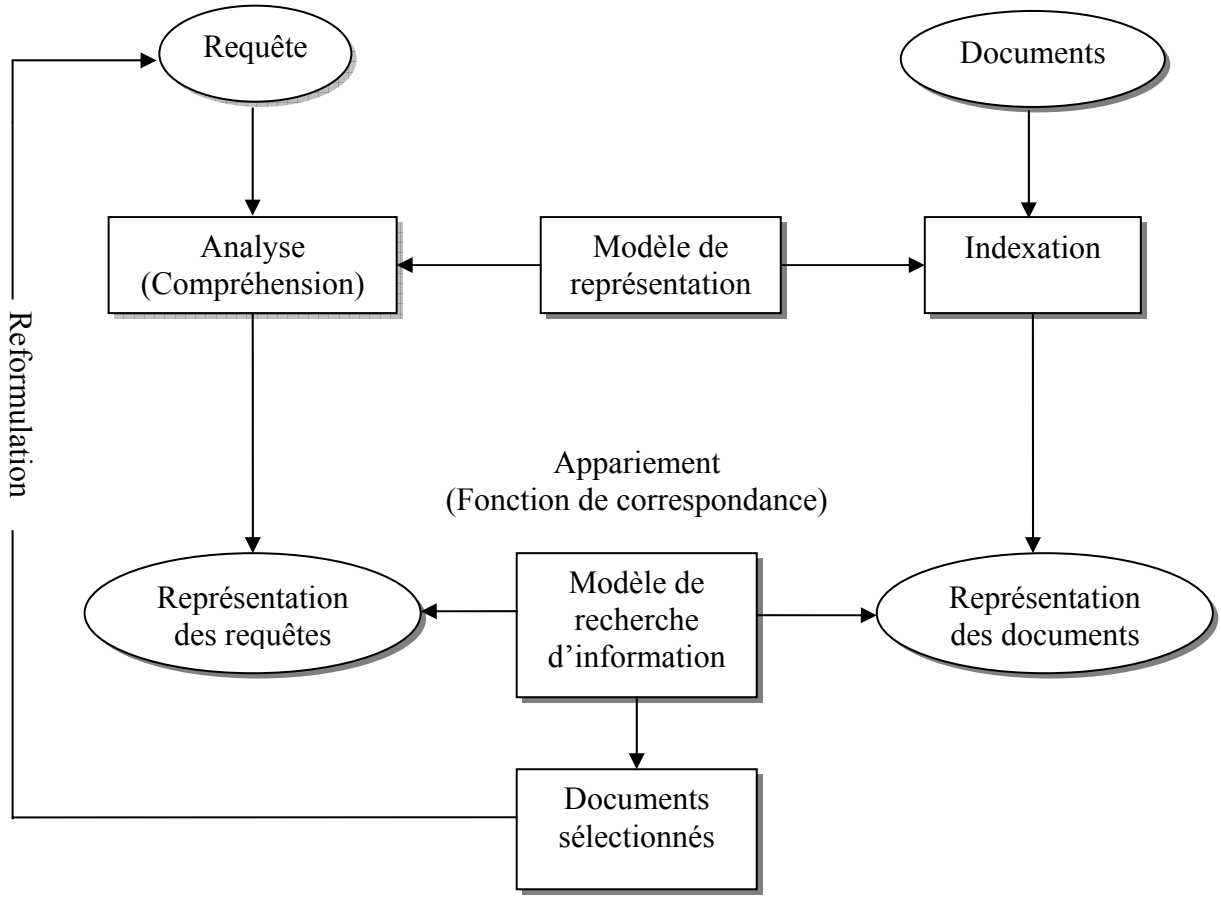


Figure 1.1 : processus en U de recherche d'information

Du schéma précédent, on peut dire que le processus de recherche d'information à partir d'une requête donnée se compose de deux processus :

- Modèle de représentation : le premier est un processus de représentation du contenu des textes (le texte étant à la fois les documents et les requêtes). Un modèle de représentation est un processus permettant d'extraire d'un document ou d'une requête, une représentation paramétrée qui couvre au mieux son contenu sémantique. Ce processus de conversion est appelé **indexation.**
- Modèle de recherche d'information : Le second est un processus de comparaison entre les représentations des textes, issues du premier processus.

1.4. Modèle de représentation :

Le but du premier processus est de représenter les documents et les requêtes, dans le même espace de représentation à l'aide d'une structure de données. Or les documents et les requêtes peuvent avoir des caractéristiques bien différentes. Par exemple, une requête peut être constituée de deux mots reliés par un opérateur booléen tandis qu'un document peut être un article de vingt pages, paru dans une revue scientifique. Donc, lorsque la différence structurelle entre les documents et les requêtes est trop importante, le processus de représentation des textes est composé en deux processus distinct appelés fonctions d'indexation : la fonction d'indexation traitant des requêtes

formulées dans un langage d'interrogation et la fonction d'indexation traitant les documents.
De manière générale, l'indexation peut être considérée comme un processus de représentation de textes. En effet, certains SRI acceptent comme requête un document entier. Dans certains cas, le meilleur document retrouvé par une requête est envoyé comme requête au SRI. Cette méthode de modification de la requête par des documents préalablement jugés pertinents, fait partie de l'approche de bouclage de pertinence [SALT, 90][Roussey, 01]. C'est pourquoi nous avons fait le choix, dans la suite de ce chapitre, de considérer l'indexation comme le processus de traitement des textes qu'ils soient documents ou requêtes.
Le but général de l'indexation est *d'identifier l'information contenue dans tout texte et de la représenter au moyen d'un ensemble d'entités appelé index pour faciliter la comparaison entre la représentation d'un document et d'une requête. Plus exactement, le processus d'indexation est le transfert de l'information contenue dans le texte vers un autre espace de représentation traitable par un système informatique* [Roussey, 01].

Tout d'abord, il nous faut définir l'espace d'indexation ou espace de représentation de l'information, en choisissant :

- Les **entités d'indexation**, qui définissent l'unité de base de l'espace d'indexation.
- La « structure » assemblant des entités d'indexation pour construire un index, c'est-à-dire une représentation.

Ensuite, il faut définir les techniques intellectuelles ou automatiques permettant, à partir du texte, de détecter les entités et de construire les structures d'indexation.

1.4.1. Les entités d'indexation :

Le résultat de l'indexation constitue le **descripteur** du document ou requête.
Les descripteurs représentent l'information atomique d'index. Ils sont censés indiquer de quoi parle le document [Laporte, 00]. On parle aussi d'unité élémentaire (tokens) [Jac&Zwe, 00]. Le but étant de les choisir de manière à ce que l'index (qui réduit la représentation) perdre le moins d'information sémantique possible.
Habituellement les descripteurs sont :

- Les **mots** du document : toute chaîne de caractères compris entre deux séparateurs (espace, caractère de ponctuation, …..).
- Le **terme, groupes de termes (concept)** : il s'agit d'expression (pouvant contenir un ou plusieurs mots). L'unité lexicale est un élément du vocabulaire de la langue, auquel sont associées des règles syntaxiques de construction de phrase. Nous définissons donc un terme comme une unité lexicale correspondant à une unité sémantique. Le terme dénote une notion précise, il est la manifestation linguistique d'un concept dans un texte [Bourigault, 98][Roussey, 01]. Autrement dit, un terme est le label d'un concept dans un contexte précis. Ces concepts peuvent être écrits de manière libre par un utilisateur ou, ce qui est souvent le cas, doivent être choisis parmi une liste de concepts (on parle alors de vocabulaire contrôlé). Cette liste de concepts sera souvent décrite dans un thésaurus (dans le cas des termes, on parlera de terminologie).

Plus rarement :

- Les **N-grames** : il s'agit d'une représentation originale d'un texte en séquences de N caractères consécutifs. On trouve des utilisations de bigrammes et trigrammes dans la recherche documentaires (ils permettent de reconnaître des mots de manière approximative et ainsi de corriger des flexions de mots ou même des fautes de frappe ou d'orthographe). Ils sont aussi fréquemment utilisés dans la reconnaissance de la langue d'un texte [Harbeck, 99].
- Les **contextes** : dans le cas du ' Lantent Semantic Indexing' [Deerwester, 90], les documents et leurs mots sont représentés sur d'autres dimensions où les mots apparaissant dans un même contexte sont proches. Cette indexation est le résultat d'une analyse des co-occurrences des mots dans un corpus.

Mais une entité d'indexation peut être tout ensemble de symboles (un nombre, une icône) caractérisant un groupe de mots (ou un groupe de termes ou un groupe de concepts, etc..), jugé valide pour représenter le contenu du document.

1.4.2. Les langages d'indexation :

L'ensemble des termes reconnus par le SRI est rangé dans une structure appelé dictionnaire constituant le **langage d'indexation**. Ce type de langage garanti le rappel de documents lorsque la requête utilise dans une large mesure les termes du dictionnaire. En revanche, il y a risque important de perte d'informations lorsque la requête s'éloigne de ce vocabulaire.
Donc l'ensemble des termes d'indexation constitue le vocabulaire du langage d'indexation. Généralement, un langage se compose d'un vocabulaire et d'une syntaxe. La syntaxe définit les règles qui régissent la formation correcte des expressions associant plusieurs éléments du vocabulaire.
On distingue deux types de langage d'indexation :

- **Langage libre :** Le langage libre est un langage évolutif, proche de notre langue naturelle (LN). Son vocabulaire, l'ensemble des éléments qui composent le langage, est choisi a posteriori et n'est pas limité par un contrôle. Le vocabulaire est composé de tous les descripteurs choisis librement pour indexer les documents. Par conséquent, le vocabulaire évolue rapidement et peut contenir des termes synonymes, polysémiques, etc. ce qui entraîne des incohérences et diminue les performances du système de recherche d'information. Par exemple, des documents portant sur le même sujet peuvent être indexés par des descripteurs différents et inversement. Ainsi un document sera retrouvé pour une requête parce que son index contient les descripteurs de la requête alors qu'il ne traite pas du sujet de la requête et inversement.

- **Langage contrôlé :** Le langage contrôlé est un langage normalisé. C'est à dire que pour éviter les problèmes de polysémie et de synonymie du langage libre, une liste de termes d'indexation est définie. Cette liste, appelée liste d'autorité, pour être efficace, ne doit pas contenir de termes polysémiques ou synonymiques. Ainsi, un terme d'indexation ne possède qu'un seul sens et inversement un sens n'est associé qu'à un seul terme d'indexation. Par conséquence, l'utilisation d'un langage contrôlé devrait permettre de limiter le nombre de représentations possibles du contenu du document, si l'on ne tient pas compte de la subjectivité de l'interprétation des documents et des termes,

car un sujet donné ne peut être décrit que par un seul ensemble de termes d'indexation. Construit a priori, ce langage doit être connu avant d'indexer un document et avant de construire une requête. Pour faciliter le choix des descripteurs et appréhender rapidement le vocabulaire du langage contrôlé, l'ensemble des termes d'indexation est organisé dans un thésaurus.

Le thésaurus contient le lexique de tous les termes normalisés (masculin singulier) du langage documentaire. Les termes du thésaurus sont reliés par des relations sémantiques qui structurent le domaine de connaissance. Ces relations, au nombre de trois [Roussey, 01]:

- La relation d'équivalence regroupe les termes jugés équivalents. C'est-à-dire que le langage documentaire ne différencie ne différencie pas ces termes les uns des autres. Ces termes peuvent être synonymes ou très proches sémantiquement. Un seul terme appelé terme préféré est choisi comme terme d'indexation. Ce terme d'indexation représente le concept identifié par l'ensemble de termes en relation d'équivalence. Par exemple, dans le thésaurus Global Legal Information Network, catastrophes, Natural Disasters et Disasters sont considérés comme synonymes et Disasters est choisi comme terme d'indexation.
- La relation hiérarchique construite une hiérarchie entre les termes d'indexation, du général au particulier ou d'un tout à ses parties.
- La relation d'association lie des termes d'indexation ayant des connotations entre eux. Par exemple l'expression « voiture à essence » forme une relation d'association entre les termes d'indexation « automobile » et « essence ».

L'organisation du thésaurus permet de trouver le terme d'indexation le plus approprié pour représenter un concept. Par exemple, l'utilisateur d'un système de recherche d'information utilise un terme de son vocabulaire comme entrée dans le thésaurus et, en suivant différentes relations, trouve le terme d'indexation reconnu par le système pour composer sa requête.

Au contraire du langage libre, certains langages contrôlés sont régis par une syntaxe permettant de composer le sens des expressions associant deux descripteurs. Par exemple, une syntaxe permet de composer les descripteurs ‘’enseignement’’ et ‘’science’’ pour différencier les index des documents traitant ‘’des sciences de l'enseignement’’ et de ‘’l'enseignement des sciences’’.

1.4.3. Les approches d'indexation :

1.4.3.1. Indexation manuelle :

C'est le documentaliste qui effectue l'analyse du document, pour identifier son contenu et construire une représentation de ce contenu (choix des mots effectué par des indexeurs). Elle est basée sur un vocabulaire contrôlé :

- Lexique : liste de mots clés
- Liste hiérarchiques : de concepts et de notations (codes)
- Thésaurus : liste de mots clés + relations sémantiques entre les mots clés.
- Ontologie : liste concepts + relations entre les concepts (la notion d'Ontologie est développée dans les chapitres suivants).

Les avantages de vocabulaire contrôlé sont :

- Permet la recherche par concepts (par sujets, par thèmes), plus intéressante que la recherche par mots simples.
- Permet la classification (regroupement) de documents (par sujets, par thème).

- Fournit une terminologie standard pour indexer et rechercher les documents.

L'indexation manuelle est souvent critiquée par les inconvénients suivants :

- Indexation très coûteuse pour construire le vocabulaire et pour affecter les concepts (termes) aux documents (imaginer cette opération sur le web).
- Difficile à maintenir puisque la terminologie évolue, plusieurs termes sont rajoutés tous les jours.
- Processus humain donc subjectif qui engendre que des termes différents peuvent être affectés à un même document par des indexeurs différents.
- Les utilisateurs ne connaissent pas forcément le vocabulaire utilisé par les indexeurs.

1.4.3.2. Indexation automatique :

C'est le système de recherche d'information qui génère les indexes des documents. L'indexation assistée (supervisée) revient, le plus souvent, à faire valider ou corriger par un humain une représentation du document proposé par le système.
L'indexation automatique présente l'avantage d'une régularité du processus, car l'indexation automatique fournit toujours le même index pour le même document. Ce qui constitue une qualité du système, mais qui est différente de la justesse de l'indexation. En effet, l'indexation automatique pêche par son incapacité à interpréter un texte et son manque d'adaptation à de nouveaux vocabulaires. Il est impossible de trouver dans les documents autre chose que ce que le système peut détecter. Par exemple, si le système n'a aucune connaissance lui permettant de lever les ambiguïtés des termes, il génèrera des erreurs d'interprétation du sens ce qui entraînera des incohérences dans la base des index.
Dans l'indexation automatisée, le contenu des textes est déterminé selon deux grandes méthodes d'analyse tel que :

- **analyse linguistique** qui repose sur les techniques du TAL, elle est fondée sur la reconnaissance des mots.
- **analyse statistique** qui est fondée sur la fréquence des mots.

1.5. Type d'indexation ou de représentation : Modèles de RI :

Les représentations, c'est-à-dire les index, sont les résultats du processus d'indexation. On peut distinguer différents types de représentation. Cette partie ne se veut pas exhaustive mais donne les quatre types les plus marquants du domaine. Cette énumération est issue de [LELO, 94][Roussey, 01]. Nous ne nous attarderons pas sur les moyens à mettre en œuvre pour construire ces index.
Si c'est l'indexation qui choisit les termes pour représenter le contenu d'un document ou d'une requête, c'est au modèle de leur donner une interprétation. Etant donné un ensemble de termes issus de l'indexation, le modèle remplit les deux rôles suivants :

- Créer une représentation interne pour un document ou pour une requête basée sur ces termes ; définir une méthode de comparaison entre une représentation de document et une représentation de requête afin de déterminer leur degré de correspondance (ou similarité).
- Le modèle joue un rôle central dans la RI. C'est le modèle qui détermine le comportement clé d'un système de RI.

Il comprend la fonction de décision fondamentale qui permet d'associer à une requête, l'ensemble des documents pertinents à restituer.

Ces modèles de recherche représentent ce qui diffère le plus entre les SRI. Ils sont inspirés de concepts mathématiques afin de pouvoir évaluer certaines relations, notamment la relation d'appariement entre les termes et les documents, qui permet au système d'obtenir une valeur de pertinence pour chaque document de la base à partir de laquelle il sélectionne ou non ce dernier.
Dans ce qui suit, on décrit quelques modèles souvent utilisés dans la RI.

1.5.1. Indexation à plat du modèle booléen :

L'indexation dite à plat considère que les descripteurs ont tout le même statut vis-à-vis du texte à indexer et qu'ils entretiennent les mêmes rapports entre eux. La représentation du texte est une succession d'entités d'indexation : une liste de descripteurs non ordonnée.
Seuls les systèmes documentaires de type booléen manipulent ce genre d'indexation.

- Modèle booléen :

C'est le premier modèle de RI basé sur la théorie des ensembles.
Dans ce modèle, un document est représenté comme une conjonction logique de termes (non pondérés), par exemple :

$$d = t_1 \wedge t_2 \wedge \ldots . \wedge t_n$$

Une requête est une expression logique quelconque de termes. On peut utiliser les opérateurs et ($\wedge$), ou ($\vee$) et non ($\neg$).
Par exemple : $q = (t_1 \wedge t_2) \vee (t_3 \wedge \neg t_4)$
Pour qu'un document corresponde à une requête, il faut que l'implication suivante soit valide :
$d \Rightarrow q$
Donc appariement exact basé sur la présence ou l'absence des termes de la requête dans les documents : Appariement (q, d) = 1 ou 0 [Boughanem, 05].
On peut remarquer les inconvénients suivants :

- La sélection d'un document est basée sur une décision binaire.
- Pas d'ordre pour les documents sélectionnés.
- Formulation de la requête difficile pas toujours évidente pour beaucoup d'utilisateurs.
- Problème de collections volumineuses : le nombre de documents retournés peut être considérable.

1.5.2. Indexation pondérée des modèles vectoriel et probabiliste :

L'indexation pondérée permet de donner à chaque descripteur un niveau d'importance. Ainsi, on peut connaître le sujet principal d'un texte et les thèmes secondaires abordés. Un poids, affecté aux descripteurs, indique son niveau d'importance par rapport au texte indexé. Ce poids est généralement calculé à l'aide d'une fonction statistique provenant d'un SRI de type vectoriel ou, dans un SRI de type probabiliste, il peut correspondre à la probabilité que le descripteur soit pertinent pour le document.

- Modèle Vectoriel :

Proposé par Salton dans le système SMART [Salton, 70][Boughanem, 05]. Dans ce modèle, un document, ainsi une requête, est représenté comme un vecteur de poids. Chaque poids dans le vecteur désigne l'importance d'un terme correspondant dans ce document ou dans la requête. Pour qu'un vecteur prenne une signification, il faut d'abord définir un espace vectoriel. L'espace vectoriel est définit par l'ensemble de termes que le système a rencontré durant l'indexation. Soit l'espace vectoriel suivant : $< t_1, t_2, \ldots, t_n >$.
Un document et une requête peuvent être représenté comme suit :

$d = < a_1, a_2, \ , a_n >$
$q = < b_1, b_2, \ , b_n >$

Où a_i et b_i correspondent aux poids du terme t_i dans le document et dans la requête respectivement.
Donc, l'ensemble des coordonnées des vecteurs est contenu dans une matrice. La fonction de comparaison évalue la correspondance entre deux vecteurs (document et requête) ce qui permet de classer les résultats.

Exemple : espacer vectoriel = $T = <t_1, t_2, t_3>$
$D_1 = <a_{11}, a_{12}, a_{13}>$
$D_2 = <a_{21}, a_{22}, a_{23}>$
La matrice représentant ce corpus de deux documents s'appelle « matrice terme-document » et s'écrit de la manière suivante :

$$\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \end{array} \begin{array}{c} D_1 \\ \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \end{pmatrix} \end{array} \begin{array}{c} D_2 \\ \begin{pmatrix} a_{21} \\ a_{22} \\ a_{23} \end{pmatrix} \end{array}$$

Etant donné ces deux vecteurs, leur degré de correspondance est déterminé par leur *similarité.*
Il y a plusieurs façon de calculer la similarité entre deux vecteurs. En voici une :

$$\text{Forme de cosinus : } sim(d,q) = \sum_i (a_i * b_i) / [\sum_i (a_i)^2 * \sum_i (b_i)^2]^{1/2}$$

Ce modèle à pour avantages :

- La pondération améliore les résultats de recherche.
- La mesure de similarité permet d'ordonner les documents selon leur pertinence vis-à-vis de la requête.

Ainsi l'inconvénient majeur est que la représentation vectorielle suppose l'indépendance entre termes.

- Modèle Probabiliste :

Le modèle de recherche probabiliste utilise un modèle mathématique fond é sur la théorie de la probabilité [Robertson, 76]. La représentation des documents est généralement un index pondéré, les poids des descripteurs correspondent à la probabilité que le descripteur soit pertinent pour le document, aussi appelée degré de croyance. Il est fondé sur l'estimation de la probabilité de pertinence d'un document par rapport à une requête que l'on exprime par P_Q(pertinence / document).

En pratique, il n'y a pas moyen d'obtenir la valeur de cette probabilité, et les systèmes s'accommodent d'une estimation de cette valeur. Pour cela, le théorème de Bayes pour des distributions discrètes appliquées au calcul de P permet d'écrire:
P(Pertinence/Document) = [P(Document / pertinence) * P(Pertinence)] / P(Document)

L'estimation de cette valeur revient à calculer :

P(Pertinence/Document) = [Pr(Document) * P(Pertinence)] / [Pr(Document)* P(Pertinence)+ Pn(Document)* P(non Pertinence)]

Où Pr(Document) et Pn((Document) représente respectivement la probabilité que le document soit pertinent ou non pertinent, P(Pertinence) et P(non Pertinence) représentent la probabilité de pertinence ou non pertinence d'un document quelconque. Ces deux derniers termes sont fixés à priori, pour un corpus donné. L'estimation des deux autres paramètres reste difficile à calculer. Dans [SM, 83], est proposée une estimation de ces valeurs basée sur le calcul d'apparition de chaque terme d'indexation dans des ensembles de documents estimés pertinents ou non pertinents.
Le modèle de probabiliste a l'avantage sur le modèle vectoriel de prendre en compte la dépendance entre les termes dans le calcul de la pertinence. Par contre, le compromis entre sa complexité de mise en œuvre et le calcul d'une estimation de la valeur théorique correcte, en fait un modèle, peu aisé à mettre en œuvre.
Une particularité de ce modèle est qu'ils ne tiennent que très imparfaitement compte du contenu sémantique des documents et des requêtes. A l'opposé, les modèles intelligents tentent d'en tenir compte en intégrant des connaissances sur le sens des termes d'indexation et leurs liaisons sémantiques.

1.5.3. Indexation structurée et le modèle logique :

Ce type d'indexation correspond à une indexation en langage contrôlé muni d'une syntaxe. Suivant la complexité de la syntaxe du langage on peut parler d'une indexation à rôle ou d'une indexation structurée.
L'indexation à rôle établit des relations sémantiques entre les descripteurs pour organiser le thème du texte à indexer. L'indexation à rôle ou à facette consiste à effectuer un rôle aux descripteurs pour définir les relations qui les associent. Les rôles possibles doivent être déterminés en fonction du domaine considéré.
Par exemple, indexons deux textes sur les activités et les services des entreprises, le premier traitant de la production de laine pour vêtements, le second de la production de vêtements en laine. L'indexation à rôle nous donne deux représentations différentes à partir des quatre rôles : ACTION, OBJET (de l'action), COMPOSITION (de l'objet), BUT (de l'action).

Production de vêtements en laine

ACTION= *production* (OBJET=*vêtement*, COMPOSITION= *laine*)

La représentation signifie qu'il s'agit d'une action particulière (produire)qui s'applique à un objet (vêtement) dont la composition est qualifiée (en laine).

Production de laine pour vêtements

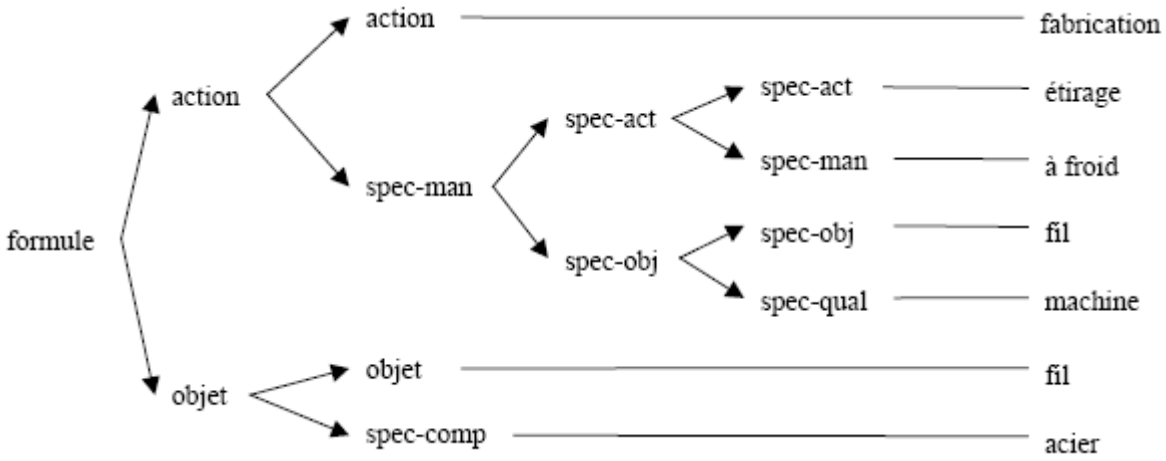
ACTION= *production* (OBJET= *laine*, COMPOSITION= *vêtement*)

La représentation signifie qu'il s'agit d'une action particulière (produire)qui s'applique à un objet (laine) pour un but particulier (vêtement).

L'indexation structurée généralise l'indexation à rôle pour lui permettre d'intégrer de nouvelles propriétés. En effet, elle permet :

- de définir des niveaux d'importance,
- d'utiliser de nombreux rôles pour définir des contextes,
- de regrouper certains descripteurs pour former des unités sémantiques (groupe de descripteurs ayant un sens à part entière),
- d'établir des relations de dépendance entre les unités sémantiques.

La représentation résultante de ce type d'indexation peut être une structure issue d'un formalisme de représentation des connaissances. L'exemple suivant montre une représentation arborescente de la phrase : ''fabrication de fil d'acier par étirage à froid des fils machine''.



La racine de l'arbre est le nœud de l'ensemble de la formule d'indexation qui se divise en action et en objet sur lequel porte l'action. Il s'agit donc ici de «fabrication de fil d'acier ». Mais l'intérêt de ce type d'index est de montrer que l'action (*fabrication*) a une spécificité de manière (*spec-man*), à savoir qu'il s'agit de *fabrication par étirage à froid*. Cette spécificité de manière est elle-même décomposée en un nom d'action qui est l'*étirage*, et dans le cas présent, elle est déterminée par le fait d'avoir lieu à *froid*. La représentation résultante de cette indexation a un pouvoir de description très riche.

Ce genre de représentation ne peut être construit qu'après une identification des concepts choisis comme entités d'indexation. Ainsi, l'ensemble de ces concepts joue le rôle de la liste d'autorité d'un langage contrôle. Dans le cadre d'une indexation automatique, une base de connaissances est nécessaire pour identifier les concepts du document à l'aide d'une analyse linguistique. Ensuite, une analyse syntaxique précise de la phrase ou du groupe de termes permet de générer une structure correcte d'association des concepts. L'assignation des rôles à partir de la structure syntaxique est opérée par plusieurs stratégies comme dans système expert [PULG, 95][Roussey, 01]. Ces systèmes utilisent des techniques de traitement automatique des langages naturelles (TALN) pour générer des structures sémantiques de représentation du texte. Par exemple, le système RECII [RASS, 94][Roussey, 01] génère automatiquement des graphes conceptuels à partir de textes médicaux. Seul un système à base de connaissances, propre à un domaine précis, et adapté à un corpus de documents homogènes, permet de générer automatiquement ce genre d'index. La lourdeur et la complexité de la construction automatique de ces représentations expliquent le fait que ces structures sémantiques ne soient pas exploitées pour des grands corpus de documents. C'est pourquoi, ce genre de représentation est plus adapté à une

indexation humaine, car la capacité ‘’d’interprétation’’ d’un système expert est loin d’égaler celle d’un documentaliste [Roussey, 01].
Ce type de représentation nécessite une fonction de comparaison adaptée, capable de prendre en, compte les différents niveaux de connaissances de l’indexation.

Actuellement, les systèmes de recherche d’information manipulent une indexation structurée se classent dans le modèle logique.

- Modèle Logique :

D’après [Roussey, 01], Van Rijsbergen [Rijsbergen, 86] modélise la pertinence d’un document répondant à une requête par une implication logique. Soit $\chi(d)$ l’information contenue dans le document d et $\chi(q)$ le besoin d’information formulée par la requête q, tous deux sont des formules logiques. Ce genre de système cherche à évaluer l’ajout minimal d’information nécessaire pour obtenir l’implication $\chi(d) \rightarrow \chi(q)$, permettant de classer les documents résultats. Cette approche améliore l’utilisation des connaissances dans le SRI car les éléments d’information et non plus les termes sont les descripteurs du document. Le problème majeur est d’extraire les éléments d’information automatiquement. La proposition de Van Rijsbergen a été appliquée à plusieurs théories logiques pour déterminer $\chi(d) \rightarrow \chi(q)$. Les théories utilisées sont la logique du premier ordre, logiques modales, logiques terminologiques. Une des théories, souvent utilisée est la théorie des situations.
Les modèles logiques développés à partir de la théorie des situations considèrent que:

- Un document est identifié à une situation,
- Les éléments d’information sont des types. Un type possède la valeur vraie dans certaine situation, et fausse dans une autre situation. La phase d’indexation détermine les types vrais dans la situation d’un document, nous avons donc affaire à une indexation à rôle.
- Des contraintes sont définies entre ces types, provenant par exemple de relations sémantiques trouvées dans un thésaurus. Ces contraintes définissent la nature du flot d’information existant entre deux situations.
- La formule de comparaison évalue l’incertitude du flot d’information circulant entre la situation du document et celle de la requête.

Les modèles logiques développés actuellement ont permis de mieux comprendre fondamentalement la recherche d’information en donnant un cadre théorique pour la comparaison entre les modèles existants. Par contre, l’implémentation de ces modèles semble difficile du fait de leur complexité c'est-à-dire qu’on fait de la RI et pas de l’intelligence artificielle : quoi intégrer, quoi négliger, comment faire des systèmes réels qui fonctionnent sur de gros corpus.

1.5.4. Indexation sémantique: Indexation basée sur les connaissances

L’indexation à base de connaissances regroupe tous types d’indexation dont le but n’est plus d’identifier l’information contenue dans le document, mais de caractériser les connaissances associées au document. Cette approche est basée sur des formalismes de représentation des connaissances comme les réseaux sémantiques et les graphes conceptuels.
L’indexation sémantique repose sur l’intuition suivant laquelle le sens des informations textuelles (et des mots qui composent les documents) dépend des relations conceptuelles entre les objets du monde auxquels elles font référence plutôt

que des relations linguistiques et contextuelles trouvées dans leur contenu [Haav, 01][Nathalie, 05]. Elle n'est possible que par l'existence et l'utilisation de ressources décrivant explicitement l'information correspondant aux objets. C'est à dire que cette approche vise à s'appuyer sur des ontologies pour représenter les documents.

1.6. Méthodes d'évaluation :

On distingue habituellement deux critères d'évaluation différents :

- Les critères quantitatifs : combien de documents peuvent être indexés, quel est le temps de réponse maximum à une requête ?
- Les critères qualitatifs : quelle est la pertinence des réponses ?

Les critères quantitatifs ne posent pas de problème, ils sont directement quantifiables : nombre de Kilo-octets maximum autorisé, nombre de secondes ou millisecondes pour la réponse à une question de taille moyenne ou le rapport temps de réponse en fonction du nombre de documents.
L'évaluation de la qualité d'un système de recherche d'information, peut se faire selon des critères subjectifs (réponse pertinente, peu pertinentes, inadaptées, aberrantes), mais, si l'on veut être objectif possible, il faut trouver des métriques pour quantifier la pertinence des réponses [Fluhr, 00].
Le principe consiste à répartir les documents en trois ensembles pour une requête donnée :

1- Les documents correspondant réellement à la requête (pertinents),
2- Les documents ne correspondant pas réellement à la requête (non pertinents), Ces deux premiers ensembles étant disjoints.
3- Les documents retournés par le SRI (retrouvés).

Ces trois ensembles peuvent être schématisés ainsi (Figure 1.2) :

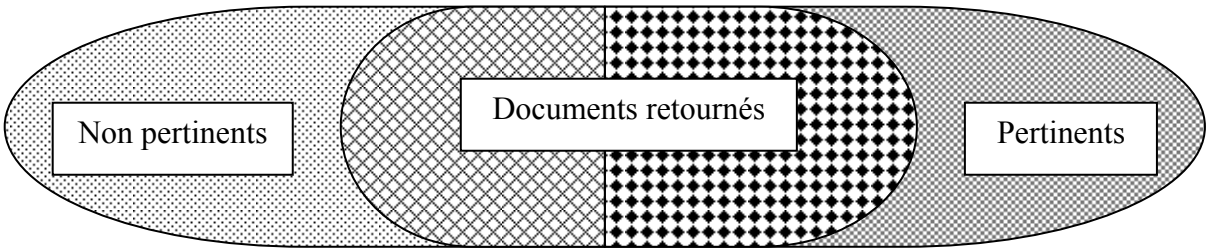


Figure 1.2 : Evaluation, classification des documents

On peut ensuite quantifier les résultats en termes de bruit et de silence (Figure 1.3) :

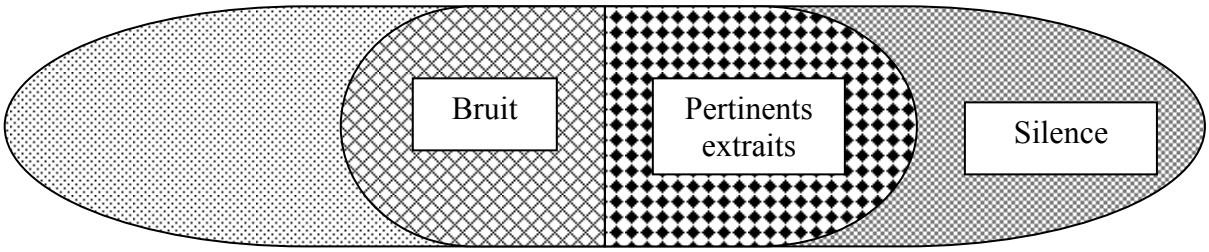


Figure 1.3 : Evaluation, bruit et silence

Le bruit représente les documents extraits mais non pertinents.

$$\text{Bruit} = \frac{\text{Nombre de documents retournés et non pertinents}}{\text{Nombre de documents extraits}}$$

Le silence représente les documents pertinents non extraits

$$\text{Silence} = \frac{\text{Nombre de documents retournés et pertinents}}{\text{Nombre de documents pertinents}}$$

Un SRI sera d'autant meilleur que le bruit et le silence seront faibles. On pourra représenter le bruit et le silence en proportion du nombre de documents extraits.
Deux autres mesures sont souvent utilisées, qui sont en fait les compléments des deux précédentes :
La précision qui représente le nombre de documents pertinents extraits par rapport au nombre de documents extraits.

$$\text{Précision} = \frac{\text{Nombre de documents retournés et pertinents}}{\text{Nombre de documents extraits}}$$

Le rappel, qui représente le nombre de documents pertinents extraits par rapport au nombre de documents pertinents.

$$\text{Rappel} = \frac{\text{Nombre de documents retournés et pertinents}}{\text{Nombre de documents pertinents}}$$

Un SRI sera d'autant meilleur que la précision et le rappel seront forts.

1.7. Critères d'une bonne indexation :

Les résultats des SRI dépendent fortement de la qualité de leur processus d'indexation. Celle-ci est jugée sur deux critères : la cohérence et l'adéquation entre les représentations des requêtes et de documents.

1.7.1. La cohérence :

Tout d'abord, une bonne indexation doit être cohérente, c'est-à-dire que deux textes traitant du même sujet, sans utiliser le même vocabulaire, sont indexés avec les mêmes descripteurs. Il est à noter que deux personnes différentes ont moins de 20% de chance de choisir spontanément le même terme pour décrire un objet [Roussey, 01]. Donc, dans le cas de l'indexation humaine, il est difficilement envisageable d'obtenir une indexation cohérente d'une grande base, sans un minimum de contrôle [FURN, 87][Roussey, 01]. D'un autre côté, l'indexation automatique n'est pas non plus cohérente car les systèmes travaillent sur les mots du document, et ne sont pas forcément capables de prendre en compte les ambiguïtés des termes comme la synonymie, ou la polysémie. Dans le cas de systèmes complexes manipulant une base de connaissances, la cohérence entre leurs indexations dépend grandement de la

qualité et de la complétude de leurs connaissances. Par exemple, si un terme n'est pas reconnu par le système il ne sera pas pris en compte pour l'indexation.

1.7.2. L'adéquation entre les représentations :

Par ailleurs, l'indexation doit vérifier un critère d'adéquation entre les représentations de la requête et du corpus. Si nous voulons retrouver un document, il faut que ses descripteurs appartiennent au même vocabulaire que ceux utilisés pour décrire la requête. Il semble plus facile de vérifier les critères de cohérence et d'adéquation en indexant avec un langage contrôlé. Par contre, celui-ci ne prend pas en compte l'évolution du vocabulaire de la base. Ce critère est le point faible des SRI indexant automatiquement, car le vocabulaire utilisé par l'auteur du document a de grande chance d'être moins actuel que celui de l'utilisateur du SRI. Il est d'ailleurs conseillé à l'utilisateur de SRI de se conformer au vocabulaire des documents plutôt qu'au sien, quant il compose sa requête. Heureusement, les procédures de bouclage de pertinence permettent à l'utilisateur d'effectuer en partie cette tâche.

1.8. Conclusion : Vers l'utilisation des techniques de TALN :

Nous avons vu dans ce chapitre les principaux concepts de la recherche d'information. Nous avons présenté l'architecture commune à tous les systèmes de recherche d'information permettant l'appariement entre les requêtes des utilisateurs et les documents de base. Notre but premier est de rappeler ce qu'est l'indexation dans ces systèmes, cette dernière que nous avons détaillée devient un processus très important pour l'élaboration de plusieurs autres applications. Puis nous avons présenté les différents modèles et stratégies utilisés lors de la mise en œuvre de ces concepts. Finalement, nous avons vu la méthode d'évaluation des SRI en terme de rappel / précision ainsi que les critères d'une bonne indexation.

Comme nous avons cité dans le paragraphe qui concerne l'indexation structurée, dans le cadre d'une indexation automatique, une base de connaissances est nécessaire pour identifier les concepts du document à l'aide d'une analyse linguistique. Ensuite, une analyse syntaxique précise de la phrase ou du groupe de termes permet de générer une structure correcte d'association des concepts. L'assignation des rôles à partir de la structure syntaxique est opérée par plusieurs stratégies comme dans système expert. Ces systèmes utilisent des techniques de traitement automatique des langages naturelles (TALN) pour générer des structures sémantiques de représentation du texte.

D'où l'indexation automatisée (dans de nombreux moteurs de recherche) repose sur les techniques de **TALN : Traitement Automatique du Langage Naturel.**

Dans la partie suivante, nous présentons la notion de **TALN.**

2. Traitement Automatique des Langues (TAL) et RI :

2.1. Introduction :

L'objectif d'un système de recherche d'information (SRI) est de retrouver parmi une masse volumineuse de documents ceux qui répondent précisément au besoin d'un utilisateur besoin formulé par le biais d'une requête en langage naturel. La principale difficulté pour ces SRI est d'établir une correspondance entre l'information recherchée et l'ensemble des documents d'une collection. Pour y parvenir, ils se fondent généralement sur un appariement entre les mots contenus dans la requête et ceux potentiellement pondérés qui représentent le contenu de chaque document. La pertinence d'un document est alors évaluée en fonction des termes communs qu'il possède avec la requête.
Compte tenu de ce mécanisme de mise en correspondance basé sur une simple comparaison de chaînes de caractères les SRI se trouvent rapidement confrontés à deux problèmes.
Le premier concerne les formulations différentes d'un même concept : un document pertinent peut contenir des termes « sémantiquement proches » de ceux de la requête mais toutefois différents (synonymes, hyperonymes, termes ayant une forme morphologique différente..).
Ce phénomène provoque une baisse du rappel de ces systèmes qui ne peuvent proposer à l'utilisateur certains documents pourtant intéressants. A ce problème vient s'ajouter celui de la polysémie des mots. L'ambiguïté qui en découle est à l'origine d'une baisse de précision des systèmes puisqu'elle entra ne potentiellement la récupération de documents non pertinents.
Pour faire face à ces difficultés liées à la complexité du langage naturel une solution souvent évoquée est d'intégrer au sein des SRI une analyse linguistique qui présente l'avantage de ne plus considérer les mots comme de simples chaînes de caractères mais comme des entités linguistiques à part entière. Les traitements linguistiques en RI, effectués par le biais de techniques du traitement automatique des langues (TAL), extraient automatiquement des informations linguistiques des documents et des requêtes. Ces connaissances ont pour ambition de permettre aux SRI une meilleure compréhension des contenus et par conséquent d'avoir un impact sur leurs performances. Les traitements linguistiques peuvent intervenir de différentes façons dans un SRI. Ils contribuent d'une part en exploitant les connaissances linguistiques extraites des textes, à améliorer le processus d'indexation des documents et, à créer une représentation plus riche de leur contenu ; cette représentation vise à obtenir un appariement plus pertinent entre l'information recherchée par l'utilisateur et les documents de la collection. Ils ont d'autre part pour objectif d'améliorer le processus de recherche des systèmes en enrichissant les requêtes par des informations complémentaires, leur offrant ainsi la possibilité de retrouver davantage de documents intéressants.
Notre objectif dans ce chapitre est de présenter une synthèse des contributions possibles des techniques issues du TAL pour une application de RI à travers un tour d'horizon des diverses tentatives qui ont déjà été réalisées dans ce domaine. En TAL on distingue généralement trois principaux niveaux d'analyse linguistique : les niveaux morphologiques, syntaxique et sémantique.

2.2. Définition :

Le traitement automatique des langues (TAL) a pour objectif de traiter des *données linguistiques (textes*) exprimées dans une *langue* dite *"naturelle"* [Delafosse, 99].

L'objectif des traitements automatiques des langues est la conception de logiciels ou programmes, capables de *traiter* de façon *automatique* des *données linguistiques*, c'est-à-dire des données exprimées dans une *langue* (dite "naturelle"). Ces données linguistiques peuvent être des *textes écrit* ou encore des *unités linguistiques* de taille inférieure à ce que l'on appelle habituellement des textes (par exemple : des phrases, des énoncés, des groupes de mots ou simplement des mots isolés).

Le traitement, dit *automatique* utilise un ordinateur c'est-à-dire une machine conçue pour effectuer des calculs. Un traitement automatique est une suite d'actions ou calculs à faire effectuer par la machine dans un certain ordre chronologique, c'est-à-dire un programme. Traiter un objet linguistique de façon automatique, implique un certain nombre de *contraintes* dans la description même de cet objet : il faut pouvoir arriver à formuler de façon totalement *explicite* et *cohérente* des ensembles de règles caractérisant le fonctionnement du texte.

Le traitement automatique des langues s'intéresse donc aux traitements informatisés qui mettent en jeu du matériel linguistique : analyse de texte, traduction automatique, etc. L'objectif est la représentation des données textuelles à différents niveaux de compréhension (morphologique, syntaxique,...)

La *recherche d'information* (RI), vise à retrouver des documents textuels répondant à un besoin informationnel, spécifié par une requête.
Donc, La recherche d'information, dans la mesure où elle travaille aussi sur des textes, s'apparente au TAL.

2.3. Grands domaines du TAL :

Nous brossons ici les grands domaines du TAL, en nous appuyant sur un découpage méthodologique classique dans le domaine et en linguistique :

☞ **La morphologie (reconnaître) :** concerne l'étude de la formation des mots et de leurs variations de forme ;

☞ **La syntaxe** (**structurer**) s'intéresse à l'agencement des mots et à leurs relations structurelles dans un énoncé ;

☞ **La sémantique (comprendre) :** se consacre au sens des énoncés ;

☞ **La pragmatique (contextualiser)** prend en compte le contexte d'énonciation.

2.3.1. Morphologie :

D'un point de vue informatique, un texte est une chaîne de caractères.
La première étape de l'analyse d'un texte est la reconnaissance, dans cette chaîne de caractères, d'unités linguistiques de base, les mots, et la mobilisation des informations associées, puisées dans un lexique.

Pour commencer, la chaîne de caractères d'entrée doit utiliser un encodage déterminé, les caractères de contrôle (fin de ligne, ;, :,etc.) étant eux aussi normalisés. On élimine généralement les caractères non répertoriés.
Il s'agit ensuite de segmenter la chaîne d'entrée en unités élémentaires (en anglais, *tokens*). Différents choix peuvent être effectués à ce stade, selon les séparateurs choisis : tous les caractères non alphabétiques (espaces, apostrophes, tirets...) ou les espaces seulement ; et selon que l'on prend en considération les « mots composés » (*« pomme de terre »* = une unité) ou pas. En tout état de cause, on est généralement amené à distinguer la notion d'unité minimale (« token ») et celle de mot (associé à une information lexicale).

Le *lexique*, en première approximation, est la liste des mots de la langue, et associe à chaque mot les informations linguistiques correspondantes : catégorie syntaxique, traits morphosyntaxiques (genre, nombre, etc.), etc. Plusieurs phénomènes amènent à préciser cette définition du lexique.

- Un mot peut exister sous plusieurs formes :
En français par exemple, formes fléchies des noms, adjectifs, etc., conjugaison des verbes. On peut alors considérer une *forme canonique* ou lemme, pour chaque mot, qui sert d'entrée dans le lexique pour l'ensemble de ses formes fléchies (singulier pour le nom, masculin singulier pour l'adjectif, infinitif pour le verbe).
L'analyse lexicale consiste à ramener les mots à une forme de base, et à reconnaître toutes les variations liées à cette forme. Trois types de formes de base :

- le radical : mang(e) pour manger, mangeoire, mangeables...
- la racine : nation pour nationalité
- le lemme : Infinitif des verbes, masculin singulier...

la lemmatisation permet de diminuer fortement le nombre de mots analysés, en éliminant toutes les flexions et les dérivations grammaticales. Son objectif est ramener chaque terme à une forme unique.

- Un mot peut avoir plusieurs sens (Polysémie) :
« avocat », *« coup »*, *« livre »* en sont des exemples ; Selon le cas, plusieurs entrées ou sous entrées sont alors distinguées.

- Plusieurs mots peuvent se trouver partager une forme commune (Homographes) :
« montre » est une forme du nom *« montre »* aussi bien que du verbe *« montrer »* ; *« pu »* est le participe passé du verbe *« pouvoir »* mais aussi de *« paître »*.

- Un mot peut être construit à partir d'un autre (par dérivation) :
(exemple : *penser, pensable, impensable)*. Ou par composition (exemple : *compter* + *gouttes = compte-gouttes).*

Enfin, pour de multiples raisons, tous les mots possibles d'une langue ne sont ou ne peuvent être répertoriés a priori dans un lexique. D'une part, les noms propres constituent un inventaire ouvert. D'autre part, de nouveaux mots sont créés régulièrement (néologie) par dérivation et composition, mais aussi par siglaison, abréviation, emprunt, etc.

2.3.2. La Syntaxe :

Pour repérer quels mots fonctionnent ensemble dans une phrase, un premier niveau de modélisation consiste à constituer des classes de mots (catégories syntaxiques, parties du discours) possédant un fonctionnement similaire : Nom (N), Verbe (V), Adjectif (A), etc.
Certaines unités, peuvent être *ambiguës* entre plusieurs catégories (ambiguïté catégorielle ou lexicale). Par exemple, chacune des unités de la phrase *« La coronarographie est normale. »* est ambiguë, ce que l'on peut noter :
« La/DET,N,PRO coronarographie/N,V est/A,N,V normale/A,N. » On remarquera que dans le contexte de la phrase entière, aucune de ces unités n'est ambiguë.
Les relations syntaxiques entre les mots d'une phrase peuvent se représenter de plusieurs façons. Le modèle en constituants considère des groupes de mots ou syntagmes, généralement centrés sur un mot de tête (N, V, etc.), et les modélise par des catégories spécifiques (syntagme nominal ou SN, syntagme verbal ou SV, syntagme adjectival ou SA, etc.). Ces syntagmes peuvent eux-mêmes être éléments d'autres syntagmes, et la structure d'une phrase est alors un *arbre de constituants* (figure 1.4.a) [Christian, 00].
Le modèle en dépendance considère directement les mots de tête (recteurs ou régissants), et leur attache les mots qui en dépendent (régis). La structure d'une phrase est alors un arbre de dépendance (figure 1.4.b) [Christian, 00]. Des équivalences existent entre les deux modèles.
Même sans ambiguïté lexicale, une phrase peut donner lieu à plusieurs structures syntaxiques (ambiguïté structurelle).

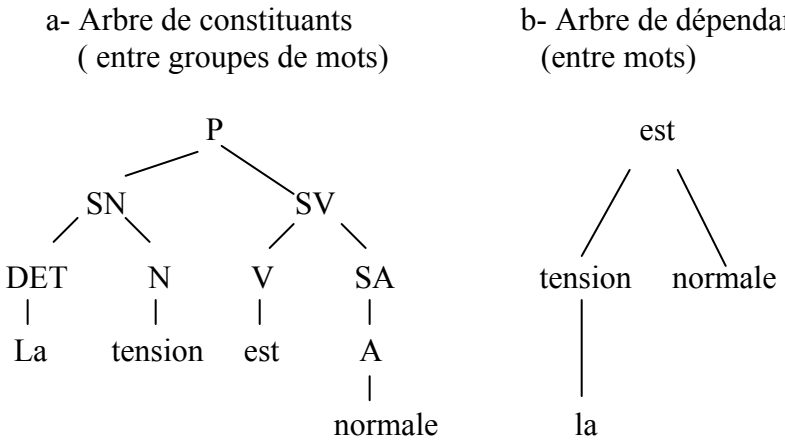


Figure 1.4 : Représentation syntaxique d'une phrase

Des relations plus précises entre mots ou syntagmes sont utiles à l'interprétation des phrases. Les relations grammaticales classiques (sujet-verbe, verbe-objet, verbe-objet-indirect, nom-modifieur, etc.) permettent de représenter la fonction des groupes de mots les uns par rapport aux autres.

2.3.3. La sémantique :

Au sens littéral, la sémantique vise à l'étude du sens hors contexte.

Le niveau sémantique est encore beaucoup plus complexe à décrire et à formaliser que les niveaux de traitements précédents, par conséquent les réalisations qui sont

opérationnelles sont peu nombreuses, et elles concernent des applications très **limitées** où l'analyse sémantique se réduit à un domaine parfaitement circonscrit ; par contre, on est encore loin de savoir construire en grandeur réelle des analyseurs sémantiques **généraux** qui couvriraient la totalité de la langue et seraient indépendants d'un domaine d'application particulier [Delafosse, 99].

Le traitement sémantique prend comme unité d'analyse la **phrase**, et conduit à représenter sa partie significative. Ces phrases, dont l'analyseur sémantique doit décrire le sens, se composent d'un certain nombre de **mots** identifiés par l'analyse morphologique, et regroupés en **structures** par l'analyse syntaxique. Ces mots et ces structures constituent autant d'**indices** pour le calcul du sens : on pourrait dire, que le sens résulte de la double donnée du sens des mots et du sens des relations entre mots [Delafosse, 99].

Donc, de manière générale, De même que pour la syntaxe, un premier niveau de modélisation consiste à constituer des classes de mots (*catégories sémantiques*). Ces classes regroupent des mots dont le sens est proche ou au minimum (pour des classes générales) des mots qui possèdent certaines propriétés sémantiques communes.
Cependant, si en syntaxe on arrive à s'accorder sur des jeux de catégories relativement consensuels, en sémantique aucune classification universelle n'existe (la constitution d'une classification universelle risque même d'être théoriquement impossible). Les classifications que l'on pourra utiliser (par exemple, les catégories générales de WordNet) reflètent nécessairement un point de vue, une prise de position culturelle ou ontologique spécifique.

Un mot, même syntaxiquement non ambigu, pourra posséder plusieurs sens. Par exemple on pourra distinguer livre qui est un verbe de livre qui est un ouvrage.

Le contexte permet en général de déterminer quel sens est à l'œuvre dans un énoncé.

Les mots d'une langue entretiennent un réseau riche de *relations sémantiques paradigmatiques* : hyperonymie / hyponymie (*« vaisseau »*/*« artère »*), méronymie (partie d'un tout : *« vaisseau »* / *« système cardiovasculaire »*), antonymie (*« malin »* / *« bénin »*) et autres contraires, etc. ce domaine de TAL se consacre au sens des textes.
Dans un énoncé, les relations grammaticales sont le support de *relations sémantiques syntagmatiques*. Par exemple, les différents actants d'un événement jouent différents *rôles thématiques* : agent, thème, source, destination, etc. Ainsi, dans *« Jean donne un livre à Marie. »*, les rôles par rapport à l'événement *« donne »* pourront être :
« Jean/agent, source donne un livre/thème à Marie/destination. »
Un mot qui désigne un événement possède des propriétés combinatoires restreintes : il sélectionne comme actants certains types de mots (*restrictions de sélection*). Ces types restreints peuvent être exprimés en termes de classes sémantiques. On pourra par exemple poser pour le verbe *« donner quelque chose à quelqu'un » donner(animé, objet, animé)* ou encore pour *« interdire » interdire(animé, animé, événement)*.

La représentation sémantique finale que l'on vise à associer à un énoncé dans un système de TAL dépend de l'objectif de ce système. Cet objectif peut être, par exemple, l'extraction d'informations spécifiques ou encor l'extraction des connaissances.

Les formalismes de représentation employés dans la représentation sémantique sont en général issus de l'Intelligence artificielle, comme les logiques de description et les Graphes Conceptuels.

2.3.4. La pragmatique :

C'est le niveau de la connaissance du monde réel et la compréhension d'un mot ou d'une phrase dans son contexte d'énonciation.

La pragmatique vise à l'étude du sens en contexte.

L'analyse sémantique de la phrase isolée, traitée hors contexte, ne conduit à représenter que la partie de la signification des mots dans cette phrase, elle n'épuise donc pas ce que l'on peut appeler la **signification complète** d'un texte, telle que l'humain l'appréhende lors d'un processus de compréhension. C'est la raison pour laquelle une analyse **pragmatique** est nécessaire, et qui consiste à trouver la signification "réelle" des phrases liées aux conditions **situationnelles** et **contextuelles** d'utilisation des mots [Delafosse, 99]. Donc, L'interprétation d'un énoncé dépend de son contexte. Dès que l'on veut traiter plus d'une phrase (et même pour une seule phrase), cette dimension intervient.
Le *co-texte* désigne le texte qui précède (et suit) la phrase courante. Deux facteurs concourent à faire qu'une phrase s'insère bien dans un texte.
– La *cohésion* régit la continuité du texte. Elle est assurée par l'emploi d'anaphores (paragraphe 2.5.3.2), l'homogénéité du thème, un emploi judicieux d'ellipses, etc.
– La *cohérence* détermine l'intelligibilité du texte. Elle s'appuie sur des structures de discours ainsi que sur les relations causales, temporelles, etc., entre les événements décrits.
Au-delà du texte lui-même, les conditions d'énonciation et les connaissances partagées complètent le contexte d'un énoncé. L'interprétation devra donc faire appel à des connaissances sur le monde. L'identification de structures de discours (structure de dialogue, structure argumentative, etc.) est également nécessaire selon le type de texte. De façon générale, une représentation de la situation décrite par un énoncé demande d'effectuer des inférences à partir de représentations initiales (par exemple, « littérales ») de cet énoncé et de représentations du contexte [Christian, 00].

2.4. Quelques pièges du langage naturel :

Le principal défi de la recherche d'information réside dans les pièges et difficultés du langage naturel. On les représente dans le tableau suivant d'après le cours [TALN, 05]:

Tableau récapitulatif
(d'après [P. Lefèvre, 00])

Caractéristiques du langage naturel	***Les difficultés dans la recherche d'informations***	***Définitions***	***Exemples***
1/ L'implicite	**La pragmatique :** Impossible à prendre en compte par des logiciels ou des langages documentaires	Liée au contexte du message, aux connaissances sur le monde, à l'usage... Domaine de la pragmatique : étude du " langage en action "	
2/ La redondance	**La synonymie :**	Mots ou expressions différents ayant le même sens ou des sens voisins.	Voiture et automobile ; tremblement de terre et séisme ; train et chemin de fer…
	La paraphrase :	Expressions équivalentes mais de structure ou de termes différents	*Mon fils a cessé de fumer* *Jean a renoncé au tabac*
	Le glissement de sens :	La **dénotation** : sens propre d'un mot La **connotation** : sens d'un mot dans un contexte particulier	*Il prend un bain* *Il est dans le bain*
3/ L'ambiguïté	**L'homonymie :**	Mots ayant la même forme, la même graphie mais des sens différents.	*Je porte la porte* *Les poules du couvent couvent*
	La polysémie :	Mots ou expressions ayant plusieurs sens	Mémoire humaine, mémoire d'ordinateur, le mémoire de maîtrise

2.5. Techniques de TAL pour la recherche d'information :

Nous examinons plus particulièrement les techniques suivantes, qui ont un impact actuel ou attendu sur la recherche d'information. Au palier morphologique, la segmentation en unités linguistiques (section 2.5.1.1) et la « racinisation » (section 2.5.1.2). Au palier syntaxique, l'étiquetage (ou désambiguïsation) syntaxique (section 2.5.2.1), l'analyse syntaxique « surfacique » (section 2.5.2.2), l'indexation sur des syntagmes (section 2.5.2.3) et la reconnaissance d'entités nommées (section 2.5.2.4). Aux paliers sémantique et pragmatique, l'étiquetage (ou désambiguïsation) sémantique (section 2.5.3.1) et la résolution d'anaphores (section 2.5.3.2). Enfin, deux techniques transversales : la statistique textuelle (section 2.5.4.1) ainsi que la traduction automatique et la recherche d'information interlangue (section 2.5.4.2).

2.5.1. Palier morphologique :

L'analyse morphologique est la première étape de traitement du LN et le préalable à toute indexation automatisée (linguistique et statistique) [TALN, 05].

Elle peut parfois constituer le seul niveau d'indexation Fondée sur l'analyse morphologique des mots : leur **forme**.

2.5.1.1. Segmentation en unités linguistiques :

Nous laisserons de côté la segmentation en paragraphes, et nous nous concentrerons sur la segmentation en phrases et en mots.
Le découpage d'un texte en phrases se fait selon les « ponctuations fortes » *«*. ! ?*»*(augmentées éventuellement du point-virgule et des deux-points). Le problème essentiel est celui de l'ambiguïté du point, qui s'utilise également pour marquer une abréviation (et aussi, en anglais, dans les nombres décimaux) : *« Le voyage aux U.S.A. de J. M. G. Le Clézio. »*. Pour réduire cette ambiguïté, on observera qu'une phrase commence par une majuscule ; le fait qu'une phrase puisse commencer par un nombre (*« 1998 a été une bonne année »*) doit aussi être pris en compte. Enfin, il faut également gérer correctement les incises qui peuvent elles-mêmes constituer des phrases ((: : :) « : : : »). Ces contraintes peuvent s'exprimer à l'aide d'automates à états finis.
La difficulté de la segmentation en mots vient du fait que les unités élémentaires (« tokens ») que l'on peut reconnaître avec sûreté ne correspondent pas toujours aux mots. Une méthode progressive consiste à segmenter dans un premier temps de façon excessivement fine (par exemple, jusqu'à 7 unités dans *« c'est-à-dire »*). Les mots contractés peuvent être eux aussi décomposés (*« du » à « de le »*, *« des » à « de les »*, etc.). Dans un second temps, on recompose les unités ainsi obtenues pour identifier des mots. On se fonde pour cela sur le contenu du lexique ou sur des modèles de mots (automates à états finis). Ainsi, la séquence *« c ' est - à - dire »* pourra être identifiée comme un mot, ainsi que *« pomme de terre »*. Cependant, certaines de ces recompositions peuvent être ambiguës : par exemple, dans *« pomme de terre cuite »*, a-t-on affaire à de la *« terre cuite »*? Ces ambiguïtés sont éventuellement propagées aux étapes suivantes de l'analyse.
En recherche d'information, la segmentation en mots est l'étape de base de l'indexation. La pertinence des unités choisies pour l'indexation influence directement la pertinence des résultats de la recherche. Une segmentation en phrases est utile pour les systèmes de résumé par extraction de phrases et pour les systèmes de question-réponse, dans lesquels les réponses fournies sont des phrases.

2.5.1.2. Racinisation :

La racinisation est une procédure plus ou moins linguistiquement fondée qui vise à regrouper les mots sémantiquement proches à partir de ressemblances « graphiques » (mots de forme apparentée). Sont généralement regroupés les mots d'un même *paradigme flexionnel* (par exemple les formes conjuguées d'un verbe avec son infinitif), et les mots d'une même *famille dérivationnelle* (par exemple un adjectif avec le substantif associé, comme *« lent »*/*« lenteur »*). En recherche d'information, la racinisation des documents et des requêtes vise à améliorer le rappel. La difficulté de

l'opération provient de la complexité et de l'irrégularité plus ou moins grande du système
Morphologique de la langue étudiée.
La racinisation peut se faire par approximation des phénomènes linguistiques en jeu ou en recherchant une fidélité linguistique plus grande. Dans la première classe de méthodes figurent les deux algorithmes classiquement utilisés en recherche d'information. Ces algorithmes ont principalement deux fonctions :
désuffixer : supprimer les suffixes qui différencient les flexions d'un mot (par exemple les formes conjuguées d'un verbe) et les mots d'une même famille morphologique (par exemple un verbe comme *« lacer »* et la forme nominale associée comme *« laçage »*),
recoder : regrouper les différentes variantes graphiques d'une même racine (ses allomorphes) comme *« condui-re »* et *« conduct-eur »*. L'algorithme de Lovins [Judith, 68][Christian, 00] effectue séparément *désuffixage* puis *recodage*, et l'algorithme de Porter [Porter, 80][Christian, 00] effectue simultanément ces deux opérations. Ces algorithmes ne sont pas exempts d'erreurs, mais donnent des résultats satisfaisants en RI pour l'Anglais par exemple.
Dans la seconde classe de méthodes se trouve la racinisation par règles et exceptions [Fiametta , 00][Christian, 00]. Chaque suffixe productif est traité par une règle (par exemple, *« -èrent »* marque les verbes du 3ème groupe au passé simple). Les formes pour lesquelles la règle ne s'applique pas (par exemple, *« légifèrent »*) ou celles, très rares, qui sont ambiguës (*« lac-èrent »* / *« lacèr-ent »*), sont listées comme des exceptions à la règle. Comme pour les algorithmes de racinisation sur l'Anglais, les *allomorphes* (mots dont les formes fléchies sont bâties sur plusieurs racines) sont réduits à une racine unique (*« cèd- »* à *« céd- »*). Cette approche permet de refléter très fidèlement les propriétés morphologiques du français.

2.5.2. Palier syntaxique :

Passage **de la forme à la syntaxe.**

Analyse syntaxique d'un texte, par un logiciel d'indexation automatique, va permettre plusieurs choses d'après [TALN, 05]:

- identification des groupes nominaux, des expressions : "accident du travail", pomme de terre", seront indexées comme expressions, et non mot par mot

- analyse syntaxique concerne la **place des mots dans une phrase**

- **reconnaissance des expressions contiguës** ou disjointes : par exemple, pouvoir reconnaître dans l'expression : Agence française de presse l'expression Agence de presse

- l'élimination des problèmes d'homographie : termes ayant la même orthographe mais de sens différent : différence entre le substantif "porte" et la forme du verbe "porte"

Donc, la syntaxe est une partie de la grammaire qui traite de la construction de la phrase. La syntaxe vise à l'étude des contraintes entre les catégories morpho-syntaxiques devant être prises en compte pour la description des séquences de mots "acceptables" dans une langue donnée. La description des contraintes caractéristiques d'une langue se fait par le biais d'une grammaire.

2.5.2.1. Etiquetage ou désambiguïsation syntaxique :

L'*étiquetage syntaxique* ou désambiguïsation syntaxique, vise à associer à chaque mot, en contexte, une « étiquette » syntaxique. Cette étiquette indique la catégorie syntaxique et éventuellement les traits morphosyntaxiques du mot. Par exemple, *« La/DETfs coronarographie/Nfs est/V3spi normale/Afs. »* L'étiquetage syntaxique est une étape intermédiaire de nombre de systèmes d'analyse surfacique ou partielle (voir plus bas), c'est pourquoi nous le présentons ici.
La plupart des méthodes cherchent à obtenir cet étiquetage en examinant le contexte immédiat du mot à étiqueter (quelques mots à gauche et à droite). Les méthodes *à base de règles* appliquent aux mots ambigus des règles de désambiguïsation, qui (selon la méthode) interdisent ou autorisent sélectivement certaines séquences d'étiquettes [Jean et al, 95][Max,00][Christian, 00]. Les méthodes probabilistes apprennent des modèles de Markov cachés sur des corpus préalablement étiquetés [Ralph, 93][Christian, 00]. La méthode de Brill [Brill, 92][Brill, 95][Christian, 00] apprend sur un corpus étiqueté des règles de correction d'erreurs d'étiquetage. Enfin, l'application d'un véritable analyseur syntaxique sur une phrase a pour effet de bord de désambiguïser les mots de la phrase [Jacque, 99][Christian, 00]. Ce dernier type de méthode n'est réellement utile dans le contexte de l'étiquetage que si l'analyse syntaxique appliquée n'a pas une complexité trop grande.
Le choix des étiquettes, et en particulier leur finesse, conditionne les performances des étiqueteurs, qui atteignent 90–98 % de mots bien étiquetés selon le jeu de catégories, le corpus, etc. La taille limitée du contexte examiné pour effectuer la désambiguïsation place une limite théorique sur la précision de l'étiquetage effectué [Jacque, 98][Christian, 00]. Par ailleurs, la plupart des mots peuvent changer de catégorie syntaxique (*conversion* d'un adjectif en nom, etc.) ; de ce fait, il est difficile de supposer que toutes les catégories syntaxiques possibles d'un mot se trouvent dans le lexique utilisé.

2.5.2.2. Analyse « peu profonde » ou « surfacique » :

Depuis le milieu des années 1980, le modèle d'analyse syntaxique dominant, fondé sur l'emploi de formalismes grammaticaux élaborés et d'analyseurs mettant en oeuvre ces formalismes, a été sérieusement concurrencé dans l'analyse de grands documents par des méthodes d'analyse simplifiées.
Ces méthodes, au moins en première intention, visent des analyses moins « profondes » ou moins complètes que les précédentes. L'*analyse partielle* ne cherche pas à traiter l'ensemble d'une phrase, mais seulement à analyser certains segments utiles et potentiellement plus faciles à reconnaître (syntagmes nominaux, syntagmes non récursifs et autres « chunks » [Steven, 91][Christian, 00]). Une méthode souvent employée est l'identification de patrons syntaxiques (typiquement, automates à états finis) dans des textes préalablement étiquetés (voir section 2.5.2.1).
Une stratégie d'*analyse robuste* fait en sorte de toujours donner un résultat, même incomplet, pour l'analyse d'une phrase. Les analyseurs « classiques » peuvent en général se replier sur une analyse partielle lorsqu'une analyse complète n'est pas obtenue (par exemple, avec les méthodes « tabulaires »). Les analyseurs qui identifient progressivement des segments de phrases « sûrs » (syntagmes non récursifs puis éventuellement syntagmes plus complexes) et les relations entre ces segments sont par nature robustes.

Enfin, l'identification de *cooccurrences* (statistiques), obtenues en recherchant des mots se retrouvant fréquemment conjointement dans une fenêtre, un paragraphe, un document, peut constituer un substitut de l'analyse syntaxique pour détecter des syntagmes élémentaires.

2.5.2.3. Indexation sur les syntagmes et variation :

Une fois que l'on a identifié des syntagmes, on peut s'en servir pour indexer les documents dans lesquels ils apparaissent. L'*indexation sur les syntagmes* (« phrase indexing ») a pour but d'augmenter la précision des index en diminuant leur ambiguïté. L'identification des cooccurrences est utilisée en RI pour faire de l'indexation sur des groupes de mots sans avoir recours à des techniques symboliques de TAL plus coûteuses à mettre en oeuvre. En concurrence, on trouve des techniques d'analyse robuste et superficielle en TAL appliquées à l'indexation pour la RI ([Fathi, 82],[Joel,87],[Koster et al,97], [Koster et al,98])[Christian, 00]. Ces techniques doivent être capables de regrouper les variantes d'un syntagme de base qui peut être modifié ou *transformé* pour produire des syntagmes de sens proche. Il est utile de savoir reconnaître ces variations pour pouvoir apparier une requête qui contient l'une des formes avec un document qui en contient une variante. Par exemple, à partir du syntagme de base *« diffusion de la lumière »*, on repérera *« diffusions de la lumière »*, *« diffusion dépolarisée de la lumière »*, *« diffusion de - lumière »*, *« diffuse une lumière »* et *« émission de lumière »*. Ces variantes peuvent être obtenues par génération dynamique de patrons de variantes (par exemple, à l'aide de métarègles) ou par simplification des structures syntaxiques des termes observés.
Parmi les enjeux de la reconnaissance de variantes, on peut citer la difficulté à couvrir exactement les variantes pertinentes et le coût computationnel de la production contrôlée de ces variantes.

2.5.2.4. Reconnaissance des entités nommées :

La notion d'*entités nommées*, introduites dans le cadre de l'extraction d'information, se réfère à des concepts uniques et partagés. Les entités nommées comprennent les organisations (entreprises, administrations, musées, etc.), les lieux (villes, régions, fleuves, etc.), les personnes (hommes politiques, vedettes, chefs d'entreprise, inconnus, etc.) et les numériques (poids, longueurs, valeurs monétaires, pourcentages, etc.). Les entités nommées peuvent constituer des index très discriminants, et sont souvent des informations demandées. Par exemple, plusieurs entités nommées sont en jeu pour répondre à la question *« Quel était le nom du PDG de Peugeot en 1987? »*.
La reconnaissance des entités nommées s'appuie sur des méthodes symboliques et numériques. Le premier type de méthode repose sur des dictionnaires (de nombreuses listes d'entités nommées sont accessibles en ligne :
Noms de lieux, annuaires divers, etc.) et des patrons syntaxiques. Ceux-ci sont appliqués sur des textes préalablement étiquetés (section 2.5.2.1) et peuvent utiliser des repères lexicaux internes (par exemple, unités pour les mesures) ou externes (par exemple, titres honorifiques pour les personnes) [David, 93][Christian, 00]. Le second type de méthode effectue un apprentissage de contextes et de structures, par exemple avec des modèles à apprentissage statistique comme les modèles de Markov cachés [Daniel, 97][Christian, 00].
On notera que l'exhaustivité des listes n'est pas l'enjeu, et que les modèles à apprentissage font mieux que les modèles symboliques. Enfin, la variation intervient

également dans l'expression des entités nommées. Abréviations et acronymes (*« MoMA » = « Museum of Modern Art »*), anaphores (section 2.5.3.2: pronoms, reprises partielles), variantes graphiques (*« ATT » = « A T T » = « AT&T » = « A T and T »*) et linguistiques (*« Ieltsine » = « Yeltsine » = « Eltsine » = « Ieltsin » = « Yeltsin »*), métaphores (*« IBM» = « Big Blue »*, *« Premier Ministre » = « Lionel Jospin » = « Matignon »*) sont autant de sources de variation qui complexifient la tâche de reconnaissance de ces entités.

2.5.3. Paliers sémantique et pragmatique :

L'analyse sémantique est fondée sur le sens des mots (les concepts). Elle va s'intéresser au regroupement de termes synonymes, aux familles de termes, pour dresser un réseau des relations sémantiques dans un texte [TALN, 05].

Systèmes relevant des systèmes experts, qui intègrent un thesaurus dans l'indexation automatique des textes [Christian, 00].

La pragmatique, comme nous avons cité, vise à l'étude du sens en contexte.

2.5.3.1. Etiquetage sémantique :

De même que l'étiquetage syntaxique (section 2.5.2.1) vise à associer à chaque mot une étiquette syntaxique, l'étiquetage sémantique cherche à associer à chaque mot, en contexte, une étiquette sémantique. Cette étiquette sémantique peut être une catégorie sémantique générale (par exemple, animé, événement, mouvement, etc.) ou un sens de mot (par exemple, *« artère »*– vaisseau sanguin *vs « artère »*–avenue). Pour une partie des mots, la désambiguïsation syntaxique peut aider : on en sait davantage sur le sens de *« livre »* si l'on connaît son genre (*« un livre/Nms » vs « une livre/Nfs »*). Par ailleurs, des méthodes similaires à celles employées en syntaxe sont applicables (chaînes de Markov, etc.). Encore plus que pour les travaux en étiquetage syntaxique, le choix des étiquettes a une influence fondamentale sur la nature de la tâche. Les travaux en désambiguïsation sémantique sont relativement récents, mais possèdent une forte dynamique.

2.5.3.2. Résolution d'anaphores :

La résolution d'anaphores consiste à relier entre elles les références à une même entité au sein d'un texte. On distingue plusieurs types d'anaphore.
L'*anaphore pronominale* emploie un pronom pour faire référence à une expression antérieure : *« Le* dispositif expérimental *d'amélioration de l'hybride est rappelé.* Il *consiste principalement en des tests. »*. L'*anaphore par reprise partielle* reprend une partie de l'expression antérieure, comme dans *« ...sont réalisés grâce à une nouvelle* technique d'immobilisation d'enzyme sur électrode de verre*. La nouveauté de* cette technique *réside dans le dépôt d'un agent... »*. L'*anaphore par lien sémantique* ne reprend pas directement un mot de l'antécédent, mais un terme sémantiquement lié (ici, plus générique) : *« La* sonde thermique INRA *est une résistance de platine...* Ce capteur *peut ainsi servir à rendre compte du phénomène... »*.
De façon générale, la plupart des expressions nominales (syntagme nominal défini, pronom) sont potentiellement des anaphores et potentiellement des antécédents d'anaphores. La résolution d'anaphores requiert des informations aussi bien syntaxiques (genre, nombre) que sémantiques (relation d'hyperonymie, etc.) et

s'appuie sur des considérations pragmatiques (entités les plus saillantes au fil du texte ou « focus ») [Cruse, 86][Christian, 00].
La résolution d'anaphores est une technique dont l'apport est important dans de nombreuses applications. En extraction d'information, elle permet de garnir une structure d'information avec la référence initiale complète à une entité. Pour répondre à une question, elle permet de donner une référence complète dans la réponse. Dans le résumé automatique, elle permet de rendre cohérentes des phrases issues de segments épars. En traduction ou en compréhension, elle permet de choisir la traduction ou le sens correct d'une anaphore ambiguë.

2.5.4. Techniques transversales :

2.5.4.1. Statistiques textuelles :

Différentes mesures d'indices textuels (mots, chaînes, catégories, patrons sytaxiques, ponctuation, taille des phrases, etc.) sont utiles dans diverses tâches liées à la recherche d'information: citons le typage du corpus, l'identification de la langue, l'ajustement des méthodes d'analyse, la catégorisation, la segmentation thématique, le filtrage de l'information, etc.

2.5.4.2. Traduction automatique et RI interlangue :

La traduction automatique, jugée comme un enjeu majeur et accessible dans les années cinquante, est désormais considérée comme une tâche extrêmement complexe. Des sous-produits de la traduction automatique rendent cependant des services appréciables autour de la recherche d'information:
Les mémoires de traduction, la recherche d'information interlangue, la constitution de données lexicales multilingues et la traduction par l'exemple.

Comme nous avons cité dans la partie précédente, La recherche d'information cherche des documents répondant à un besoin informationnel ou *sujet* (figure 1.5), exprimé à l'aide d'une *requête*. Les documents sont au préalable *indexés* : Chaque mot de chaque document est répertorié dans une table inverse, avec ou sans conservation des positions des mots dans le texte d'origine. L'appariement entre la requête et l'index va déterminer les documents qui sont considérés comme répondant le mieux au besoin informationnel initial.

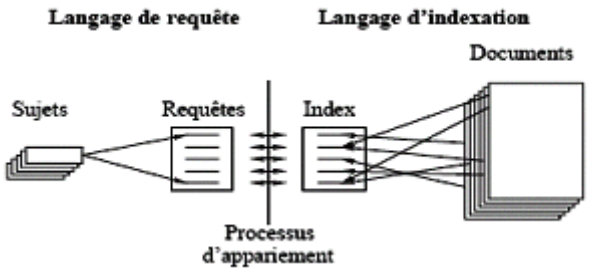


Figure 1.5 : schéma général de la recherche d'information

Et comme nous avons cité dans le chapitre précédent au niveau de paragraphe d'architecture d'un SRI, le processus de recherche d'information à partir d'une requête donnée se compose de deux processus : le modèle de représentation appelé

indexation. Et le modèle de recherche d'information qui est un processus de comparaison entre les représentations des textes, issues du premier processus.
Mais avant d'être traité le premier processus c'est à dire l'indexation, il faut faire une étape préalable qui concerne la simplification de documents et de la requête.
Cette simplification vise à rendre plus pertinent et plus efficace le processus d'appariement entre requête et index. Elle s'effectue selon les étapes suivantes :
1- Suppression des « mots stop » (mots grammaticaux, mots fréquents, mots sans pouvoir discriminateur...) ;
2- racinisation (*stemming*) (paragraphe 2.5.1.2, réduction des mots de la même famille morphologique à une racine commune),
3- transformation du texte en un sac (ou un ensemble) de mots,
4- amalgame (*conflation*) des mots synonymes.

Une extension de schéma de la figure 2.3 permet d'effectuer de la *recherche d'information interlangue* : le sujet de recherche est formulé dans une langue (par exemple, français) différente de celle des documents (par exemple, anglais).
Dans ce cas, le système de RI inclut une étape de traduction du sujet en une requête dans la langue cible. Les documents trouvés peuvent en retour être également traduits dans la langue source [Christian, 00] .
Un point clé dans la *recherche d'information interlangue*, est la traduction d'une requête dans la langue cible de la recherche. Cette tâche fournit un exemple prototypique de la façon dont un problème (la traduction de termes complexes) peut être reconsidéré sous un angle différent dans le contexte de la recherche d'information.
Un terme étant donné (par exemple, *« groupe de travail »*), il est bien connu qu'une traduction assemblée mot à mot à l'aide d'un dictionnaire bilingue (*« groupe »* → *« cluster »*, *« group »*, *« collective »*, ... ; *« travail »* → *« work »*, *« labour »*, ...) a toutes les chances d'être incorrecte, voire incongrue (*« work cluster »*, *« labour cluster »*, *« work collective »*, etc.).

Cependant, si l'on examine le nombre d'occurrences parmi les documents (du Web par exemple) des termes produits, on peut généralement identifier la ou les traductions correctes. Par exemple, la recherche sur AltaVista d'une chaîne fixe de mots et le relevé des nombres de documents contentant cette chaîne permet de déduire la ou les traductions correctes d'un terme complexe.

2.6. Conclusion :

Les techniques de TAL peuvent être intégrées à une ou plusieurs composants d'un SRI afin de fournir une compréhension de texte à différents niveaux. Les techniques d'analyse linguistiques profondes ont donc les impacts sur la performance de recherche d'un SRI, concrètement sur le taux de précision/rappel. Evidement, ils peuvent être aussi appliqués dans les travaux d'extraction de connaissances à partir de texte.

Nous développerons ci-après les méthodes d'extraction de connaissances à partir de texte.

3. Extraction des connaissances à partir des textes:

3.1. Introduction :

La première phase de processus d'indexation à base de connaissances consiste à extraire les connaissances (termes + relations) des documents à indexer.
Malgré les avantages, de cette approche, la création d'un index structuré est un processus difficile et coûteux (temps, personnes…). Ceci a amené les chercheurs à travailler sur la possibilité de (semi-)automatiser cette tâche en utilisant des techniques d'acquisition des connaissances à partir des textes.
Nous définissons pour cela deux objectifs majeurs dans le premier consiste à acquérir / extraire des termes significatifs et représentatifs du contenu informationnel du texte (des termes désignant des instances des concepts de l'ontologie). Le deuxième objectif consiste à acquérir / extraire des relations entre ces termes (des instances des relations de l'ontologie).
Ces objectifs ne sont pas très éloignés des objectifs de l'indexation automatique des documents pour les SRI classiques. En effet, la plupart de ces méthodes essayent de capturer des termes représentatifs du contenu informationnel du corpus et des relations (simples) reliant ces termes. La grande différence entre ces deux problématiques (index structuré et indexation à partir des textes) réside dans le fait que pour la première, (i) l'extraction des connaissances est guidée par un modèle déjà prédéfini du domaine (i.e. l'ontologie) et (ii) les termes extraits ne représentent pas uniquement le contenu informationnel du document mais représentent aussi les connaissances du domaine traitées dans ce document.
Les utilisations possibles des résultats sont multiples :

- Aide à la construction manuelle de ressources comme les thésaurus, les réseaux sémantiques ou les ontologies ;
- Indexation ;
- Recherche d'informations : extension de requête avec des mots sémantiquement proches ;
- Désambiguïsation sémantique.

Dans cette partie, nous présentons une rétrospective de différents travaux qui traient l'extraction de connaissances. D'abord, nous classons ces travaux en 3 classes selon l'approche utilisée : l'approche statistique, l'approche linguistique et l'approche hybride.
Ensuite, nous présentons quelques travaux sur les syntagmes et la recherche d'information.

3.2. Unités lexicales et conceptuelles :

3.2.1. Mots clés :

En recherche d'information, les mots clés sont les mots qui décrivent le mieux le contenu d'un document ou d'un corpus. Les mots clés sont souvent des noms, des verbes ou des adjectifs, par opposition aux mots outils comme les prépositions, les déterminants ou les pronoms.
En linguistique de corpus, les mots clés sont les mots qui apparaissent plus fréquemment dans un document que ne le voudrait le hasard. De nombreuses mesures

reposant sur les différences de fréquence d'occurrence permettent ainsi d'extraire automatiquement les mots-clés d'un document ou d'un corpus.

3.2.2. Termes :

Le terme est défini de la manière suivante par [Roche, 2005][Delphine , 06] :
Élément d'une terminologie. Combinaison indissociable d'un concept et d'une dénomination (désignation).
Du point de vue classique, celui de E. Wüster et du Cercle de Vienne, le terme est la dénomination d'un concept, chaque concept étant désigné de manière non ambiguë par un seul terme [Jacquemin et Bourigault, 03][Delphine , 06]. La dénomination ou désignation1 d'un terme peut prendre diverses formes linguistiques. La distinction la plus courante est faite entre **termes simples** ou **monolexicaux** et **termes complexes** ou **polylexicaux**. Dans le premier cas, le terme est composé d'un seul mot graphique, dans le second d'une succession de mots.
Le terme est donc un élément construit, dont le statut est différent des autres mots de la langue car il répond à un besoin de normalisation sémantique. De plus, on considère généralement que les termes doivent être monosémiques dans le domaine considéré.
Ainsi, les méthodes d'acquisition automatique de connaissances présentées dans les sections suivantes concernent l'acquisition de **candidats termes** qui doivent encore être validés et normalisés par les terminologues au cours d'un processus qui ira du mot ou de l'expression au concept [Rastier, 1995, Jacquemin et Bourigault, 03] [Delphine , 06].

3.2.3. Unités de sens : Concepts ou catégories conceptuelles :

Un concept est la représentation mentale d'un en semble d'objets différents, mais considérés comme équivalents d'un certain point de vue (nom identique, action commune, etc.). Les concepts ne se trouvent pas directement dans les textes. En effet, comme le constate très justement C. Roche [Roche, 2005][Delphine , 06], « Il n'y a pas de concepts dans un texte, mais uniquement des traces linguistiques de leurs usages ».

3.3. Relations sémantiques

Après avoir passé en revue les différents types d'unités décrites dans les ressources lexicosémantiques (ontologies), nous allons maintenant décrire les relations sémantiques établies entre ces unités.
Ces relations sont distribuées sur deux axes :

- **Axe syntagmatique** (horizontal). Deux mots sont en relation syntagmatique s'ils apparaissent ensemble dans un texte. On dit également que les mots sont co-occurrents s'ils apparaissent ensemble dans un contexte restreint. Les relations sémantiques définies sur l'axe syntagmatique sont de type associatif comme par exemple entre *tasse* et *café* (La *tasse* contient du *café*) ou *chat* et *lait* (le *chat* boit du *lait*).
- **Axe paradigmatique** (vertical, hiérarchique). Deux mots sont en relation paradigmatique s'ils apparaissent dans des contextes similaires. C'est à ce

niveau que l'on retrouve un certain nombre de relations structurant le lexique telles que la méronymie et l'hyponymie.

En psychologie, et notamment dans les études sur la formation des concepts et des catégories, on distingue deux types de catégories : les catégories taxonomiques et les catégories thématiques [Lin et Murphy, 2001, Nguyen et Murphy, 2003, Wisniewski et Bassok, 1999][Delphine , 06] :

- Les catégories **taxonomiques** sont organisées en hiérarchies de catégories de plus en plus abstraites comme *bouledogue* < *chien* < *mammifère* < *animal*. Ces catégories sont basées sur des propriétés communes ou la similarité.
- Les catégories **thématiques** groupent des objets qui sont associés ou qui ont une relation de complémentarité (les entités ne jouent pas le même rôle). On trouve différents types de relations thématiques : spatiale (un *toit* se trouve sur une *maison*), fonctionnelle (un morceau de *craie* sert à écrire sur un *tableau noir*), causale (l'*électricité* fait briller l'*ampoule*) et temporelle (le *repas* est suivi de la *note* au restaurant). Celles-ci sont généralement moins représentées dans les ressources lexico-sémantiques (ontologie).

Nous allons maintenant détailler les différents types de relations sémantiques pour les mots qui sont en relation paradigmatique. Ces relations sont pour la plupart d'entre elles décrites dans les thésaurus et les ontologies.

3.3.1. Relations d'inclusion et d'identité :

3.3.1.1. Synonymie :

Selon [Cruse, 00] les synonymes sont des mots dont les similarités sémantiques sont plus saillantes que les différences. Il est alors possible de distinguer différents degrés de synonymie : synonymie absolue (identité de sens, ce qui est très rare), synonymie propositionnelle (les termes peuvent se substituer l'un à l'autre dans un contexte linguistique particulier sans altérer les conditions de vérité de la phrase) et synonymie proche (comme par exemple *mist* et *fog*). Les termes synonymes correspondent au même concept. La relation de synonymie est symétrique, mais pas nécessairement transitive [Lafourcade et Prince, 01].

3.3.1.2. Hyponymie :

La relation d'hyponymie (encore appelée subsomption, spécialisation, relation ISA ou EST-UN) implique un rapport d'inclusion entre les sens des mots. Par exemple, *pomme* et *pêche* sont des **hyponymes** de *fruit* et *fruit* est un **hyperonyme** de *pomme* et de *pêche*. On dit également que *pomme* et *pêche* sont subsumés sous *fruit* et qu'ils sont tous deux **co-hyponymes** de *fruit*. La relation de spécialisation est transitive et non symétrique.

3.3.1.3. Méronymie.

La relation de méronymie (aussi appelée relation PART-OF ou PARTIEDE) correspond à la relation partie-tout. Ainsi, *globule* est un **méronyme** de *sang* et *sang* est un **holonyme** de *globule*.

3.3.2. Relations d'exclusion et d'opposition :

3.3.2.1. Co-hyponyme :

Deux termes co-hyponymes peuvent avoir des sens **incompatibles**, c'est-à-dire qu'ils ne peuvent être vrais en même temps, comme par exemple *chien*, *chat*, *souris* ou *lion* qui sont tous des hyponymes d'*animal*.

3.3.2.2. Complémentation :

Deux termes sont **complémentaires** si l'un implique le contraire de l'autre, comme par exemple *mort* et *vivant*.

3.3.2.3. Antonyme :

Deux termes qui sont des **antonymes** stricts appartiennent à la même catégorie syntaxique et sont opposés sur un axe gradué, comme par exemple *long* vs *court*, *chaud* vs *froid*, *bon* vs *mauvais*. On peut considérer l'**antonymie** comme « un cas particulier de la relation de **co-hyponymie** » [Amsili, 03][Delphine , 06]. Cette relation est notamment utilisée dansWordNet pour l'organisation des adjectifs [Hayes, 1999] [Delphine, 06].

3.4. Les approches d'extraction de termes :

On peut distinguer deux types de méthodes d'acquisition automatique de termes et de mots clés :

- Les méthodes à base de patrons définissant la structure des termes à extraire.
- Les méthodes à base de calculs statistiques (mesures d'association et de comparaison).

Ces deux types de méthodes ne s'excluent pas mutuellement et peuvent être combinés pour obtenir de meilleurs résultats.

3.4.1. Méthodes à base de patrons :

Les termes polylexicaux peuvent être caractérisés par des patrons reposant essentiellement sur l'étiquetage morpho-syntaxique. Ce pré-requis n'est toutefois par indispensable car certains systèmes, comme celui proposé par J. Vergne ou le système ANA, fonctionnent à partir de corpus de textes bruts, non étiquetés. Les patrons peuvent également être définis à partir de la structure morphologique des termes, permettant ainsi l'identification de termes simples (composés d'un seul mot graphique) morphologiquement complexes.

3.4.1.1. Patrons morpho-syntaxiques :

Cette famille de méthode nécessite deux types d'informations préalables : l'étiquetage morphosyntaxique du corpus ainsi qu'un ensemble de patrons reposant sur cet étiquetage et décrivant la structure des termes que l'on cherche à extraire. Les termes ainsi identifiés sont généralement des groupes nominaux [Kageura et Umino, 1996] [Delphine, 06].

C'est l'une des techniques les plus utilisées pour l'extraction de termes. Les systèmes basés sur cette technique supposent que les termes à extraire obéissent à des régularités syntaxiques stables. Ces systèmes prennent en entrée un ensemble de patrons constitués d'une suite de catégories grammaticales et qui peuvent être par exemple : NOM NOM / ADJQ NOM / NOM PREP NOM …
Toutes les occurrences de mots correspondant à ces patrons sont extraites comme des candidats termes potentiels. Parmi ces analyseurs on peut citer :

Nomino [David, 90][Nathalie, 05], quant à lui, repose sur le découpage des textes en unités lexicales pour l'identification de syntagmes nominaux. Il est considéré comme l'un des premiers systèmes à avoir utilisé cette technique. Proposé à la base pour la construction de bases de connaissances. NOMINO implémente toutes les étapes du traitement linguistique, il détecte ainsi les noms présents dans le document/corpus et en s'appuyant sur des règles d'expansion (une grammaire de patrons morpho-syntaxiques) propose une liste syntagmes nominaux triée, soit par fréquence, soit par ordre alphabétique.

Contrairement à NOMINO, LEXTER [Bourigault, 96] prend en entrée un corpus préalablement étiqueté et désambiguïsé. Cet outil permet également l'extraction de candidats termes sous forme de syntagmes nominaux décomposés en tête et expansion. LEXTER implémente une méthode originale qui consiste à éliminer d'abord les mots ne pouvant constituer un terme (verbe, conjonction, pronom…) pour ensuite relever des syntagmes nominaux maximaux.
Le système LEXTER développé par D. Bourigault [Jacquemin et Bourigault, 03] [Véronique, 05] est un analyseur syntaxique robuste dédié à l'extraction de syntagmes (nominaux et adjectivaux) à partir de corpus spécialisés, dans une perspective d'acquisition terminologique. Il procède en deux étapes pour extraire les unités terminologiques. Dans un premier temps, il repère les groupes nominaux maximaux en se basant sur des règles permettant d'identifier les limites de syntagmes nominaux les plus vraisemblables. Puis il décompose ces groupes nominaux afin d'en extraire les termes candidats. De plus, les termes candidats sont organisés sous forme de réseau en fonction des éléments lexicaux partagés dans des positions syntaxiques similaires.
[Bourigault et al., 05] propose une évolution de LEXTER vers un nouveau système appelé Syntex en rajoutant deux extensions importantes : (i) la prise en compte de l'anglais, et (ii) l'extension de la couverture du système à l'extraction des syntagmes verbaux.
Ces trois systèmes proposent un réseau de termes dont les relations lexicales (tête et expansions) peuvent conduire à des relations sémantiques.
Le système ACABIT (Automatic Corpus-based Aquisition of BInary Terms) [Daille, 96] a pour objectif de préparer la tâche du terminologue en lui proposant une liste ordonnée de candidats-termes pour un corpus préalablement étiqueté et lemmatisé. Les candidats-termes correspondent à un type particulier de co-occurrences où sont prises en compte les nombreuses variations des termes : variations flexionnelles et syntaxiques faibles, variation de modification interne, variation de coordination et variations attributives. Le candidat-terme présenté à l'expert est une forme générique regroupant les différentes occurrences du candidat-terme rencontré dans le corpus sous sa forme de base ou sous la forme d'une de ses variations. Les candidats termes sont classés suivant un score d'association. Cette méthode ne fait donc pas

uniquement appel à des filtres linguistiques permettant de repérer certains types de syntagmes nominaux mais utilise également des mesures statistiques.

3.4.1.2. La méthode de Jacques Vergne :

Contrairement aux systèmes que nous venons de présenter, la méthode proposée par J. Vergne ne nécessite aucun étiquetage morphosyntaxique préalable des textes. Elle permet d'extraire des termes de structure contrôlée par des patrons reposant sur l'alternance dans le texte de mots informatifs et non informatifs. Ces catégories sont définies de la manière suivante [Vergne, 05], selon la distinction introduite par Lucien Tesnière : Les mots informatifs sont les mots pleins ou lexicaux (content words), et les mots non-informatifs sont les mots vides ou grammaticaux (function words).
A cette définition linguistique, J. Vergne fait correspondre des indices facilement mesurables en corpus : « un mot informatif est plus long et moins fréquent que ses voisins », reprenant ainsi les principes de Zipf (« ce qui est d'usage fréquent est court ») et Saussure (« dans la langue, il n'y a que des différences »). La méthode ne faisant usage d'aucune ressource externe au corpus est à la fois endogène et multilingue. Elle évite donc le recours à une *stop list* (liste de mots outils), nécessairement propre à une langue, et potentiellement ambiguë car contenant des mots correspondant à des homographes qui peuvent être informatifs ou non informatifs suivant le contexte. Les différentes étapes de la méthode, dans sa version la plus récente [Vergne, 05], sont détaillées comme suit :

- **Données** Texte à indexer.
- **Étape 1** Identification des mots informatifs (mots pleins) et non informatifs (mots vides) à l'aide des différences de longueur et d'effectif entre mots contigus.
- **Étape 2** Génération de candidats termes de structure contrôlée en utilisant des patrons reposant sur l'étiquetage en mots informatifs (I) et mots non informatifs (n) : $I+$, $I + n + I+$, $I + n + I + n + I+$.
- **Étape 3** Suppression des termes hapax et des termes inclus dans des termes de même effectif.
- **Étape 4** Calcul du poids de chaque terme dans le document en fonction de son effectif et de sa longueur.
- **Résultats** Liste de termes pondérés.

3.4.1.3. Système ANA, Apprentissage Naturel Automatique d'un Réseau Sémantique :

Le système ANA [Enguehard, 92] se distingue également par l'absence de pré-traitement des corpus. Il est inspiré par l'apprentissage humain de la langue maternelle. Le programme extrait une liste initiale de termes du domaine constituant un « bootstrap », ainsi que des listes de mots fonctionnels et de connecteurs de termes complexes. Dans la phase de découverte, les termes complexes contenant un des termes de la liste de bootstrap sont repérés et leurs composants sont ajoutés à la liste de bootstrap. Ils peuvent alors être réutilisés pour la découverte de nouveaux termes, dans un processus itératif.
Les différentes étapes de la méthode sont décrites comme suit :

- **Données** Corpus de textes bruts.
- **Familiarisation** Extraction automatique de quatre listes : mots fonctionnels, mots fortement liés, mots de schémas, bootstrap (termes du domaine).

- **Découverte** Extension de la liste des termes du domaine à partir du bootstrap en utilisant les patrons suivants :
  - expression : terme constitué de deux termes co-occurrents.
  - candidat : co-occurrence d'un terme, d'un mot de schéma et d'un mot (nouveau terme).
  - expansion : co-occurrence d'un terme et d'un mot.
- **Résultats** Liste de termes.

3.4.1.4. Patrons morphologiques :

Les méthodes à base de patrons précédemment décrites définissent la structure lexicale des termes polylexicaux. Or, le vocabulaire de domaines spécifiques, comme la médecine, se caractérise également par des patrons de formation incluant des **segments** de mots spécifiques, comme par exemple le suffixe *-ite* en médecine [Heyer *et al.* 2006] [Delphine, 06]. Ces spécificités, marquées par l'utilisation d'affixes typiques, peuvent être mises à profit pour l'acquisition de termes. Pour ce faire, il est dans un premier temps nécessaire de repérer les affixes spécifiques au domaine.
[Ananiadou, 1994][Delphine, 06] propose un système d'analyse morphologique des composés savants conférant le statut de termes aux mots contenant certains suffixes et éléments de formation. Les suffixes sont identifiés manuellement à partir de l'analyse d'un corpus de spécialité. Le système décrit par [Heid, 1998] [Delphine, 06] extrait quant à lui les mots contenant certains préfixes, suffixes et éléments de formation caractéristiques de domaines techniques en allemand comme *mega+*, *mikro+*, *+gramm* ou *+graph*. Certaines listes d'affixes spécifiques à des domaines précis comme la médecine sont disponibles. Il est également possible de les identifier de manière automatique.
En effet, ces affixes sont rares dans le vocabulaire général et peuvent donc être identifiés par comparaison de leur fréquence d'occurrence dans un corpus de spécialité par rapport à un corpus de langue générale (voir Section 1.3.3, p. 23). Cette méthode permet également la découverte de radicaux spécifiques. Voici quelques exemples de morphèmes identifiés par comparaison à partir de textes légaux en allemand : *Beratung+*, *Amt+*, *Abnahme+*, *+recht*, *+gericht*, *+betrieb*, *+betrag*.
La méthode d'identification d'éléments spécifiques par comparaison des fréquences est également utilisée par [Witschel, 2005][Delphine , 06] pour la découverte de trigrammes (suites de trois lettres) spécifiques d'un domaine. Les mots qui contiennent ces trigrammes sont les termes candidats.
Cette méthode est surtout efficace pour les domaines techniques dont les termes contiennent des éléments de formation grecs ou latins1. On trouve par exemple les trigrammes spécifiques suivant, à partir d'un corpus de textes anglais sur l'asthme : *sth* (*asthma*, *anti-asthmatic*), *uco* (*glaucoma*, *mucosa*), *thm* (*asthma*, *dysrythmia*), *apy* (*therapy*, *immunotherapy*) [Heyer *et al.*, 2006][Delphine , 06] .

3.4.2. Mesures d'association (mesures statistiques)

Ces mesures permettent de quantifier l'information partagée par des couples de mots ou termes et de repérer les groupes de mots qui apparaissent ensemble plus fréquemment que ne le voudrait le hasard. Ces mesures se basent sur l'hypothèse que l'emploi de deux termes en cooccurrence est l'expression d'une relation sémantique entre ces termes [Rij79, JC94][Haddad, 02]. Cette relation s'exprime par des combinaisons de mots qui occurrent souvent dans un corpus et dont le statut

linguistique (par exemple les catégories grammaticales des mots) peut varier. Il existe beaucoup de formules différentes pour ces mesures d'association, nous n'en présenterons que quelques-unes. Leur calcul se base généralement sur des tables de contingence semblables à la Table 1.1. Cette table de contingence contient les effectifs observés *O* pour les couples de mots apparaissant dans un contexte donné (co-occurrence directe, phrase, etc.). Les effectifs sont mesurés pour les couples de mots qu'il est possible de former à partir de deux mots *x* et *y* et l'ensemble des autres mots du corpus. L'effectif observé pour le couple de mots *xy* est noté $O11$, celui du couple $\neg xy$ est $O21$, etc. La taille de contexte utilisée pour leur calcul est variable, même si pour l'acquisition de termes on ne prend généralement en compte que la co-occurrence directe (mots adjacents).
Les paragraphes suivants détaillent quelques-unes de ces méthodes.

	$Y = y$	$Y \neq y$
$X = x$	$O11$ $f(x, y)$	$O12$ $f(x, \neg y)$ $f1(x) - f(x, y)$
$X \neq x$	$O21$ $f(\neg x, y)$ $f2(y) - f(x, y)$	$O22$ $f(\neg x, \neg y)$ $N - f1(x) - f2(y) + f(x, y)$

Tab. 1: Table de contingence pour deux éléments *x* et *y*. N correspond au nombre de tokens, $f1(x)$ au nombre d'occurrences de *x* en première position dans le couple et $f2(y)$ au nombre d'occurrences de *y* en deuxième position dans le couple.

3.4.2.1. Fréquence de co-occurrence :

La mesure la plus simple pour déterminer la force d'association entre deux ou plusieurs mots est de compter le nombre d'occurrences de la suite de mots considérée dans le corpus : les suites de mots qui apparaissent fréquemment dans le corpus pourront être considérées comme des termes.
Cette approche est celle adoptée par la technique des **segments répétés** qui consiste à repérer les séquences de mots répétées dans le corpus [Lebart et Salem, 94]. L'identification des segments répétés repose sur les caractères délimiteurs comme les signes de ponctuation : les segments ne peuvent chevaucher un signe de ponctuation (délimiteur de séquence). Un segment répété est une suite d'occurrences non séparées par un délimiteur de séquence et de fréquence supérieure ou égale à 2. Donc, cette approche s'appuie sur la détection de chaînes constituées de morceaux (mots, symboles, ponctuation…) existant plusieurs fois dans le même texte [Delphine , 06]. Cette approche commence par stocker tous les mots du texte dans une table dont la valeur correspond soit à un terme, soit à une ponctuation, soit à un symbole de structure du texte (saut de paragraphe, chapitre, etc.) et une fréquence minimale d'apparition dans le texte est fixée afin d'éliminer les faibles occurrences. Pour chaque forme du texte, l'ensemble des suites dans le texte commençant par cette forme est répertorié. Le processus est réitéré pour chaque forme du texte.
Les résultats ainsi obtenus peuvent être améliorés par l'utilisation de filtres [Rousselot, 2004][Delphine , 06] identifiant les mots qui indiquent des frontières de termes, en complément des signes de ponctuation délimiteurs de séquences (filtre «

coupant » : verbes courants, adverbes, pronoms relatifs, conjonctions) et ceux qui ne peuvent se trouver aux bornes d'un terme (articles, prépositions).
Les redondances sont ensuite supprimées en utilisant le mécanisme de l'**intersection lexicale** (également appelé **contrainte d'autonomie** par [Drouin, 2003][Delphine , 06]) : par exemple, si l'on trouve *artère coronaire droite* de fréquence 3 et *coronaire droite* de fréquence 3, *coronaire droite* est considérée comme un sous-segment de *artère coronaire droite* et est donc supprimée.

3.4.2.2. Le test du χ^2 :

La formule pour le test du χ^2 (ou *test de Pearson*) est la suivante :

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

Pour le test du χ^2 on fait l'hypothèse que les variables sont distribuées de manière normale.
Cependant, cette hypothèse n'est pas très réaliste dans le cas de la fréquence d'occurrence des mots dans un texte car les événements rares sont très fréquents [Dunning, 1993][Delphine , 06].

3.4.2.3. Le coefficient de Jaccard :

La formule du coefficient de Jaccard est la suivante :

$$\text{Jaccard} = \frac{f(x, y)}{f1(x) + f2(y) - f(x, y)} = \frac{O_{11}}{O_{11} + O_{12} + O_{21}}$$

Cette mesure peut également s'appliquer à la comparaison de vecteurs de co-occurrences pour donner une mesure de la similarité entre mots [Manning et Schütze, 1999], [Oakes, 1998][Delphine , 06].

3.4.2.4. L'information mutuelle :

L'information mutuelle compare la probabilité d'observer deux mots *x* et *y* ensemble aux probabilités de les observer indépendamment. Donc, permet de détecter des cooccurrences de deux mots en comparant la probabilité de les trouver simultanément avec la probabilité de les trouver indépendamment. Cette information se calcule de la manière suivante, selon la formule donnée par [Church et Hanks, 1990][Delphine , 06] :

$$\mathbf{IM}(x, y) = \log_2 \frac{P(x, y)}{P(x)\,P(y)}$$

P(x) et P(y) sont les probabilités d'observer les mots x et y, et P(x, y) est la probabilité de les observer simultanément (sans notion d'ordre). C'est-à-dire $P(x, y)$, $P(x)$ et $P(y)$ sont estimés par le maximum de vraisemblance tel que $f(x)$ est l'effectif du mot *x* dans un corpus de taille *N* (nombre d'occurrences de mots) :

$$P(x, y) = \frac{f(x,y)}{N} \quad ; \; P(x) = \frac{f(x)}{N} \quad ; P(y) = \frac{f(y)}{N}$$

Le calcul de la fréquence de co-occurrence de x et y, $f(x, y)$ peut s'effectuer dans des fenêtres de différentes tailles. Dans le cas des bigrammes (suite de deux mots) de la table de contingence 1.1 la formule devient donc :

$$I(x, y) = log \frac{Nf(x, y)}{f_1(x)f_2(y)}$$

Selon la formule, si x et y sont liés, $P(x, y)$ sera supérieur au calcul de la fréquence de cooccurrence attendue $P(x)P(y)$, sous l'hypothèse nulle de l'indépendance des occurrences de x et de y, et donc $I(x, y) >> 0$.c'est-à-dire Si les deux mots x et y sont dépendants l'un de l'autre, l'IM a une forte valeur positive. S'ils ne sont pas en relation, la valeur est proche de 0. Si les deux mots sont en distributions complémentaires, la valeur est négative.

Cependant, l'information mutuelle ne fonctionne pas très bien pour les données d'effectif faible dans le corpus considéré [Manning et Schütze, 99][Dunning, 93][Delphine , 06]. Une autre mesure, le log du rapport de vraisemblance (*log likelihood ratio* en anglais) est plus robuste eu égard au traitement d'événements dont le nombre d'occurrences est faible. Cependant, la valeur de cette mesure est élevée pour des termes fréquents qui apparaissent rarement ensemble. Les deux mesures peuvent donc être combinées pour obtenir de meilleurs résultats [Pantel et Lin, 2001][Delphine , 06].

3.4.2.5. Coefficient de Dice :

La mesure du coefficient de dice est une mesure symétrique qui s'exprime comme suit:

$$\textbf{Dice}\ (x, y) = \frac{2\ P(x, y)}{P(x) + P(y)}$$

où P(x) et P(y) sont les probabilités d'observer les mots x et y, P(x,y) est la probabilité de les observer simultanément (sans notion d'ordre) et $0 \leq \text{Dice}\ (x, y) \leq 1$.

Si les deux mots x et y sont dépendants l'un de l'autre le Coefficient de Dice a une valeur proche de 1. Dans le cas contraire, le Coefficient de Dice a une valeur proche de 0.

Donc, ces approches utilisent seulement les co-occurrences de mots. Le principe est que si deux mots co-occurrent souvent dans un certain type de contexte, alors ils peuvent être regroupés dans un terme.

Le calcul de co-occurrences varie selon le contexte et selon les besoins. Il peut se faire dans le même document, le même paragraphe, la même phrase, ou dans une certaine distance.

3.4.2.6. Limites des mesures d'association

Outre les inconvénients propres à chacune des mesures décrites, les mesures d'associations ne permettent pas de repérer avec précision les unités terminologiques. Ces mesures sont d'une part limitées par le nombre de mots, généralement deux, auxquelles elles peuvent s'appliquer.

De plus, les frontières naturelles entre unités peuvent ne pas être respectées car elles n'utilisent aucune information sur la structure des unités à extraire.

3.4.3. Évaluation des résultats de l'extraction terminologique :

Les résultats de l'extraction des termes ou de mots clés sont difficiles à évaluer, et ce quelle que soit la méthode utilisée (patrons, mesures statistiques). Les mesures d'évaluation utilisées viennent du domaine de la recherche d'information. Elles utilisent soit une liste de référence des termes du domaine, généralement produite indépendamment du corpus utilisé par l'extraction, soit une annotation manuelle d'un ou plusieurs documents.
Le **rappel** mesure la capacité de la méthode à identifier tous les termes du document de référence. Il se calcule en divisant le nombre total de termes correctement identifiés par le nombre de termes dans le document de référence.

$$\text{Rappel} = \frac{|\ T \cap Tref\ |}{|\ Tref\ |}$$

Avec :

- $|\ T\ |$ = nombre total de termes identifiés
- $|\ Tref\ |$ = nombre de termes dans le document de référence
- $|\ T \setminus Tref\ |$ = nombre de termes correctement identifiés

La **précision** mesure la capacité de la méthode à identifier des termes corrects. Elle se calcule en divisant le nombre total de termes correctement identifiés par le nombre total de termes identifiés.

$$\text{Précision} = \frac{|\ T \cap Tref\ |}{|\ T\ |}$$

On cherche ainsi à obtenir la plus grande précision et le plus grand rappel possible. La **F-mesure** combine la précision et le rappel et correspond à leur moyenne harmonique.

$$\text{F\_mesure} = \frac{2 \cdot \text{précision} \cdot \text{rappel}}{\text{précision} + \text{rappel}}$$

Enfin, la mesure de rappel en fonction du rang des candidats ou mesure de « **ranked recall** » [Streiter *et al.*, 2003][Delphine , 06], donne une mesure de la qualité du classement des termes en fonction d'un poids. Les termes ou mots-clés extraits sont souvent pondérés par une mesure, comme la fréquence d'occurrence dans le corpus, ceci afin de les trier et donc de faciliter leur analyse. Les termes dont le poids se situe en deçà d'un certain seuil peuvent ainsi être éliminés.
Il est donc utile d'évaluer les rangs attribués aux mots-clés par de telles mesures. En effet, une méthode d'extraction qui identifie 5 termes corrects, apparaissant aux rangs 3 à 7 est moins bonne qu'une autre méthode qui identifie les même mots-clés mais qui les classes du rang 1 à 5.
Si *ri* est le rang du *i*-ème mot-clé extrait et $n = |\ T \setminus Tref\ |$ alors la mesure de « ranked-recall » se calcule comme suit :

$$\text{Ranked recall} = \frac{\sum_{i=1}^{n} i}{\sum_{i=1}^{n} r_i}$$

Malgré l'éventail de mesures qu'il est possible d'utiliser, l'évaluation des résultats de l'acquisition de termes candidats est une tâche difficile. On distingue deux types d'évaluation : utilisation d'une ressource existante ou validation manuelle des termes candidats par des experts du domaine.
La méthode d'évaluation la plus courante consiste à comparer la liste de termes candidats obtenue à une liste de référence des termes du domaine construite manuellement et validée [Daille, 96]. Cependant, les terminologies de référence existantes sont généralement établies indépendamment d'un corpus textuel précis, tandis que les résultats de l'extraction automatique sont fortement dépendants du contenu du corpus utilisé. Ainsi, tous les termes pertinents par rapport au corpus ne sont pas forcément présents dans la terminologie de référence et peuvent donc pour cette raison être injustement considérés comme du bruit.
S'il n'existe pas de terminologie de référence du domaine, reste la méthode de la validation manuelle. Celle-ci n'est toutefois possible que pour un corpus de taille restreinte ou pour un nombre réduit de termes, généralement les meilleurs selon le poids assigné par la méthode.
Ainsi, [Enguehard, 1992] présente plusieurs évaluations manuelles du système ANA, pour des listes de termes d'une taille limitée, comprise entre 300 et 700 éléments. L'évaluation proposée par [Streiter *et al.*, 2003] [Delphine , 06] se base quant à elle sur un corpus très petit de 994 mots (occurrences).
De plus, l'évaluation manuelle doit être effectuée par des spécialistes du domaine, qui pourront ne pas s'accorder sur la liste des termes corrects.
Il est également possible de combiner les deux méthodes d'évaluation. Dans [Drouin, 2004][Delphine , 06], les termes extraits (spécificités) sont d'abord comparés avec une banque de terminologie. Ceux qui sont présents dans la terminologie sont considérés comme pertinents tandis que les autres sont évalués par des terminologues afin de déterminer leur pertinence.
Dans tous les cas, c'est généralement la pertinence (pertinence par rapport au domaine et pertinence par rapport au corpus) qui est mesurée, plus que la mesure classique de rappel, difficile à estimer car il faudrait alors disposer de l'inventaire de l'ensemble des termes à extraire du corpus.
Une fois les termes, reste à les structurer. Nous allons donc maintenant décrire les systèmes d'acquisition de relations sémantiques utilisés pour organiser les termes.

3.5. Les approches d'extraction de relations sémantiques:

Il existe deux approches principales pour l'acquisition de relations sémantiques entre termes. Les méthodes dites « externes » se basent sur la comparaison des contextes d'occurrence, tandis que les méthodes dites « internes » reposent sur la structure morphologique des mots ou la structure lexicale des expressions. Nous allons dans un premier temps décrire diverses approches contextuelles, comme les modèles à base de vecteurs et de graphes, les méthodes de classification et les méthodes par description de patrons lexico-syntaxiques. Puis, nous allons présenter les méthodes reposant sur la structure interne des mots et expressions.

3.5.1. Vecteurs et graphes de co-occurrences :

De nombreux modèles d'acquisition de relations sémantiques reposent sur l'idée que le sens d'un mot est lié à ses contextes d'utilisation. Ainsi, les mots sont considérés comme sémantiquement proches s'ils apparaissent dans des contextes similaires. Ces

méthodes rejoignent ainsi les théories qui mettent l'accent sur l'importance de l'usage pour la sémantique lexicale telles que celles du linguiste Firth (« You shall know a word by the company it keeps ») et du philosophe Wittgenstein (« Meaning is use »).
Différents niveaux de relations de co-occurrence peuvent être considérés :

- **Co-occurrences de 1er ordre** ou co-occurrences directes : deux mots sont considérés comme proches s'ils apparaissent dans le même contexte, c'est-à-dire s'ils sont directement co-occurrents (relation syntagmatique).
- **Co-occurrences de 2nd ordre** ou co-occurrences indirectes : deux mots sont similaires s'ils apparaissent dans des contextes similaires (relation paradigmatique). Ainsi deux mots *M*1 et *M*2 pourront être considérés comme sémantiquement proches s'ils partagent des co-occurrents *Mi* et ce même s'ils n'apparaissent jamais dans le même contexte [Denhière et Lemaire, 2003, Martinez, 2000, Rapp, 2003][Delphine , 06].

La co-occurrence directe est mesurable de diverses manières, présentées dans la section 3.4.2 : Fréquence de co-occurrence, information mutuelle, test du χ^2 , rapport de vraisemblance, indice de Jaccard. De plus, suivant la méthode, la taille du contexte (fenêtre de mots, phrase, paragraphe) est variable. Les mots constituant le contexte sont parfois sélectionnés en fonction de leur catégorie morpho-syntaxique.
La mesure de la co-occurrence indirecte s'effectue généralement à l'aide de vecteurs représentant chaque mot. Chaque composante d'un tel vecteur contient la mesure de co-occurrence directe du mot considéré avec un certain mot du lexique. Il faut noter que certaines méthodes représentent les co-occurrences sous forme de graphe dont les noeuds sont les mots et les arcs représentent la relation de co-occurrence directe ou indirecte entre mots. Les graphes de co-occurrence sont surtout utilisés pour découvrir les sens et usages différents des mots par détection des zones de forte densité dans le graphe. Les deux modes de représentation, vecteurs et graphes, ne sont toutefois pas totalement dissimilaires, un graphe pouvant être représenté sous forme de matrice.
Les mesures basées sur les vecteurs consistent à calculer la similarité de deux mots en fonction de la similarité ou de la distance des vecteurs les représentant. Les mesures les plus fréquemment utilisées sont la distance euclidienne, la mesure de Kullback-Leibler et le cosinus. Généralement, la taille des vecteurs est réduite de sorte à ne conserver qu'un certain nombre de composantes.
En effet, le nombre de mots d'un corpus est de l'ordre des dizaines de milliers et les vecteurs résultants sont donc très grands et qui plus est « creux »(beaucoup de composantes ont pour valeur 0). Les composantes conservées sont sélectionnées en fonction de divers critères : fréquence (les mots les plus fréquents), variance (composantes de plus grande variance), productivité (composantes non nulles pour le plus grand nombre de vecteurs). De plus, le nombre de dimensions conservées est également variable1.
Les modèles vectoriels les plus connus sont LSA (Latent Semantic Analysis) et HAL (Hypertext Analog to Language). Le principe de LSA [Landauer *et al.*, 1998][Delphine , 06] consiste à transformer tout corpus en une matrice dans laquelle chaque ligne correspond à un mot et chaque colonne à un contexte. Chaque cellule de la matrice contient le nombre d'occurrences d'un mot dans le contexte correspondant à la colonne. Le contenu de la matrice est pondéré puis soumis à une décomposition en valeurs singulières pour réduire le nombre de dimensions de la matrice à environ 300. Les mots peuvent alors être comparés en calculant la similarité des lignes de la matrice qui les représentent. HAL [Lund et Burgess, 1996, Li *et al.*, 2000][Delphine ,

06] procède différemment pour construire la matrice. Les lignes contiennent les valeurs de co-occurrence pour les mots précédant le mot correspondant à la ligne dans le corpus, tandis que les colonnes représentent les mots suivants. Chaque mot peut ainsi être représenté par un vecteur dont la taille est le double de celle du lexique (concaténation du vecteur ligne et du vecteur colonne). En pratique, la taille du vecteur est réduite aux composantes présentant la plus grande variance.
Il faut également noter que les modèles à base de vecteurs peuvent être utilisés pour l'acquisition des sens des mots polysémiques. Par exemple, le système ACOM [Ji, 2004][Delphine , 06] sélectionne des mots liés par les contextes, appelés contexonymes puis forme des cliques à partir de ces mots contextuellement liés, une clique correspondant aux mots qui sont tous les contexonymes les uns des autres. Les cliques sont alors projetées dans un espace sémantique multi-dimensionnel et regroupées par un algorithme de classification hiérarchique. Ce système est basé sur la méthode initialement proposée par [Ploux et Victorri, 1998][Delphine , 06], permettant de caractériser le sens des mots polysémiques à partir de cliques construites à l'aide de dictionnaires de synonymes. Dans ce cas, les cliques sont des unités de sens, similaires aux synsets de WordNet : dans une clique, tous les mots sont synonymes (ou quasi-synonymes) aux autres mots de la clique.
Ces approches sont néanmoins critiquables sur certains points :

- Il est nécessaire de disposer de très gros corpus de textes pour obtenir de bons résultats car la plupart des mots sont rares (problème connu sous le nom de « data sparseness »).
- Ces méthodes donnent une mesure de proximité entre termes. Mais la nature exacte de la relation (i.e. une des relations décrites dans la Section 3.3) n'est pas connue. Par exemple, une étude des relations sémantiques effectivement extraites par LSA montre que seule une faible proportion de ces relations correspond à des relations d'inclusion, d'identité ou d'opposition.
- Les unités contextuelles généralement utilisées sont les mots. Or les mots ne sont pas des entités sémantiquement atomiques mais décomposables en unités porteuses de sens de niveau inférieur, les morphèmes.

Les mesures de similarité obtenues peuvent être utilisées pour catégoriser automatiquement les mots, grâce à des algorithmes de classification que nous décrivons dans la section suivante.

3.5.2. Classification :

L'objectif de la classification est de repérer les mots similaires pour ensuite les grouper en catégories.
Les algorithmes de classification de mots sont variés. Dans certains cas, la classification est hiérarchique : les catégories obtenues forment alors une taxinomie.
Les cartes auto-organisatrices de Kohonen (ou SOM pour *Self-Organizing Maps*) sont utilisées pour classer divers types de données, et notamment les documents [Kohonen *et al.*, 2000] [Delphine , 06].
Elles permettent également de catégoriser les mots. Les cartes auto-organisatrices sont un type simple de réseau de neurone dont l'objectif est d'organiser les données dans un espace à deux dimensions, la carte, et ceci de manière totalement non supervisée. [Honkela *et al.*, 1995][Delphine , 06] présentent une utilisation des SOM pour la classification des 150 mots les plus fréquents des contes de Grimm. La carte obtenue après apprentissage reflète à la fois les catégories sémantiques et les catégories syntaxiques des mots étudiés.

Les méthodes de classification utilisent généralement des textes analysés syntaxiquement, afin de repérer diverses relations permettant de former des classes de mots. Ainsi, [Caraballo, 1999][Delphine , 06] se base sur les groupes nominaux apposés ou joints par une conjonction de coordination pour construire automatiquement une hiérarchie de noms en utilisant une méthode de classification hiérarchique ascendante.
L'utilisation des contextes syntaxiques peut permettre une classification encore plus raffinée, par le regroupement dans des classes distributionnelles des termes qui apparaissent dans le même contexte syntaxique. La similarité de deux mots est alors fonction du nombre de contextes syntaxiques qu'ils partagent. Ce type d'analyse distributionnelle se fonde sur les travaux de Harris [Habert et Zweigenbaum, 2002][Delphine , 06]. Par exemple, [Hindle, 1990][Delphine , 06] décrit une méthode de classification de mots anglais en fonction des structures prédicat-argument dans lesquels ils apparaissent. En effet, un nom ne peut généralement être le sujet et/ou l'objet que d'un nombre restreint de verbes. Si l'on prend l'exemple du mot *wine*, il peut apparaître avec les verbes *drink* et *produce* mais pas *prune*. Il est ainsi possible de caractériser chaque nom par un ensemble de verbes. Puis, les noms peuvent être regroupés en fonction des similarités des environnements lexico-syntaxiques dans lesquels ils apparaissent.
La première étape du traitement consiste donc à effectuer une analyse syntaxique du corpus, afin de mettre à jour les relations de type sujet-verbe-objet. La pertinence de ces relations est évaluée à l'aide de l'information mutuelle. Les noms sont ensuite regroupés à l'aide d'une mesure de similarité basée sur l'information mutuelle calculée et prenant en compte les verbes partagés.
D'une manière assez semblable, [Lin, 1998][Delphine , 06] présente une méthode d'identification de mots similaires pour la construction automatique de thésaurus. Tout d'abord, un analyseur syntaxique est appliqué au corpus pour obtenir des couples de mots liés par une relation de dépendance comme sujet-verbe, verbe-objet, nom-adjectif, etc. Puis une mesure de similarité entre mots est calculée en fonction des relations de dépendances partagées par les mots.
Divers autres systèmes se basent sur la même procédure: analyse du corpus pour extraire des relations syntaxiques et agrégation des mots partageant les mêmes relations syntaxiques.
Cependant, même si ces méthodes se basent sur un corpus analysé et sur des relations de co-occurrences bien spécifiques, elles ne permettent pas toujours d'étiqueter les relations sémantiques obtenues. De plus, le bénéfice de l'utilisation d'une analyse syntaxique préalable n'est pas vérifié dans tous les cas. En effet, les méthodes utilisant l'analyse syntaxique fournissent de meilleurs résultats pour les mots les plus fréquents mais sont surpassés pour les méthodes utilisant une fenêtre de mots sans autre pré-traitement pour les mots les moins fréquents. De plus, le travail sur de très gros corpus de textes nécessite de prendre en compte également le temps d'exécution et la taille des représentations fournies par les analyseurs [Curran et Moens, 01][Delphine , 06].

3.5.3. Patrons lexico-syntaxiques :

Les méthodes décrites précédemment ne permettent pas l'acquisition de relations sémantiques étiquetées. Or certaines relations comme l'hyper/hyponymie se caractérisent par des constructions spécifiques qu'il est possible de repérer dans les textes après les avoir spécifiées sous forme de patrons lexico-syntaxiques.

Afin d'améliorer la pertinence des couples de termes hyponymes extraits par cette méthode, [Cederberg et Widdows, 03][Delphine , 06] proposent d'utiliser l'analyse sémantique latente (LSA) pour effectuer un filtrage. Plus la similarité des deux termes est importante suivant cette analyse, plus la relation d'hyponymie qui les relie est plausible. [Hearst, 92][Delphine , 06] décrit également une méthode permettant de découvrir de nouveaux patrons :

1. Choisir une relation lexicale pour laquelle on souhaite découvrir les patrons.
2. Réunir un ensemble de termes liés par cette relation.
3. Rechercher dans le corpus les contextes dans lesquels les couples de termes apparaissent ensemble.
4. Trouver les points communs de ces contextes. [Morin, 98] définit une mesure de similarité permettant de regrouper ces environnements dans des classes.
5. Lorsqu'un nouveau patron a été identifié, l'utiliser pour rassembler de nouvelles instances de la relation et revenir à l'étape 2.

Une variante de cette méthode consiste à la combiner avec une approche non supervisée. Dans ce cas, les relations plausibles statistiquement (selon une mesure d'association des termes telle que l'information mutuelle) sont sélectionnées de manière non supervisée, ce qui permet d'automatiser les phases 1 et 2 de l'extraction.
[Rebeyrolle, 00] complète les patrons morpho-syntaxiques par l'utilisation de marqueurs typographiques et dispositionnels pour repérer les définitions dans un texte. Ainsi, le terme défini peut être marqué typographiquement par des caractères gras ou italiques, des lettres majuscules ou des guillemets. De plus, les structures définitoires se retrouvent régulièrement en début de paragraphe. Au niveau discursif, le terme à définir est généralement mentionné une première fois avant d'être repris dans un énoncé définitoire.
Les méthodes d'acquisition de relations sémantiques à partir de textes que nous venons de présenter dans les trois sections précédentes ont toutes un point commun, celui d'utiliser le contexte d'occurrence des mots. Dans le cas le plus simple, aucun pré-traitement n'est appliqué au corpus et le contexte est alors constitué de mots. Dans les cas les plus complexes, certains patrons spécifiques, basés sur l'analyse morpho-syntaxique du corpus, sont recherchés dans le corpus. Or, les termes sont souvent des unités polylexicales. Certaines méthodes, que nous allons décrire dans la section suivante, se basent donc sur la structure lexicale des termes pour leur structuration.

3.5.4. Utilisation de la structure interne des termes :

Les informations internes aux termes, reposant sur leur structure, peuvent être utilisées, notamment pour repérer les relations d'antonymie et de spécialisation. Les informations internes utilisables se trouvent à deux niveaux :

- Niveau du morphème : la comparaison entre termes simples, monolexicaux mais polymorphémiques, s'effectue sur la base de leur structure morphologique.
- Niveau du mot : la comparaison entre termes polylexicaux s'effectue sur la base des mots qu'ils contiennent.

Nous allons d'abord présenter les méthodes basées sur la structure lexicale des termes polylexicaux, qui sont celles que l'on rencontre le plus fréquemment dans la littérature. Puis, nous présentons les utilisations possibles de la structure morphologique pour l'acquisition de relations sémantiques.

3.5.4.1. Utilisation de la structure lexicale des termes polylexicaux :

Certaines relations sémantiques, et notamment l'hyperonymie et la co-hyponymie sont marquées par des relations structurelles entre les termes.
De nombreux travaux se basent sur l'**inclusion lexicale** pour retrouver des relations d'hypo-/hyperonymie. En effet, l'hyperonymie se manifeste par une structure lexicale spécifique, notamment pour les noms : l'hyperonyme est un nom, l'hyponyme est un composé, comme par exemple *table gigogne* ou *table de cuisin*. Le nom hyperonyme est inclus dans le composé qui est son hyponyme.
La notion d'inclusion lexicale est définie de la manière suivante par [Grabar et Zweigenbaum, 2002a][Delphine , 06] : un terme *T*1 est lexicalement inclus dans un autre terme *T*2 ssi tous les mots informatifs formant *T*1 se trouvent également dans *T*2. Par exemple, le terme *acide gras* est inclus dans le terme plus long *acide gras libre* : *acide gras* est l'hyperonyme de *acide gras libre*, c'est-à-dire qu'un *acide gras libre* est un type d'*acide gras*. On peut distinguer trois types de relations d'inclusion lexicale en anglais :
– **Expansion gauche** : *T*2 = *M* + *T*1. *M* peut être selon le cas un adjectif [Bodenreider *et al.*, 2001, Ibekwe-SanJuan, 2005], comme par exemple *ventricular aneurysm – aneurysm* ou un nom, comme dans *compression fracture – fracture*.
– **Insertion** : dans ce cas, un nouveau mot *M* est inséré au milieu de *T*1 pour former *T*2 comme dans *adult brain glioblastoma – adult glioblastoma*.
– **Expansion droite** : *T*2 = *T*1+*M*, comme par exemple *cholesterol – cholesterol granuloma*.
Les relations hiérarchiques sont marquées de manière préférentielle par l'expansion gauche et l'insertion. L'expansion droite correspond à des relations sémantiques plus faibles, similaires aux liens *Voir aussi* présents dans les thésaurus.
Les résultats obtenus par les méthodes basées sur la structure lexicale montrent qu'elles complètent utilement les méthodes contextuelles et offrent l'avantage d'identifier des liens sémantiques spécifiques comme la spécialisation. Les liens identifiés pourraient donc permettre de compléter les relations hiérarchiques présentes dans la ressource, après validation.
Une autre relation structurelle, qui présente un intérêt pour l'identification de co-hyponymes, est la **substitution**. Elle correspond au remplacement d'un mot informatif de *T*1 par un autre mot dans *T*2 où *T*1 et *T*2 contiennent le même nombre de mots.
La substitution de modifieurs indique une relation de co-hyponymie entre termes, comme c'est le cas par exemple pour *liver granuloma – cardiac granuloma*.
Les modifications morpho-syntaxiques de l'un des constituants d'un terme complexe peuvent induire des relations sémantiques supplémentaires comme l'antonymie (*organic chemical – inorganic chemical*) ou la succession dans le temps [Daille, 03][Delphine , 06].

3.5.4.2. Utilisation de la structure morphologique des termes simples :

Dans la mesure où de nombreux termes techniques ont une structure morphologique complexe, il est envisageable d'appliquer des méthodes similaires aux termes simples. L'utilisation de la structure morphologique des termes simples pour l'acquisition de relations sémantiques est bien sûr dépendante des ressources morphologiques disponibles. La majorité des systèmes combinent donc acquisition de données morphologiques, de manière supervisée ou non supervisée, et utilisation des données

ainsi extraites. Nous décrirons plus précisément les méthodes d'acquisition de connaissances morphologiques dans le chapitre suivant.
Les relations sémantiques sont marquées dans la structure morphologique de diverses manières.
L'hyponymie se manifeste sous forme d'inclusion dans les mots à composition savante, comme l'atteste l'exemple suivant : « *Angiosarcome* : type de *sarcome* qui apparaît sur un vaisseau san-guin ». Ainsi, [Buitelaar et Sacaleanu, 02][Delphine , 06] présentent une méthode d'extension des synsets de GermaNet (version germanique de WordNet) reposant sur la décomposition des mots complexes comme *Sauerstofftherapie* (oxygénothérapie) en tête (*Therapie*) et modifieur (*Sauerstoff*). Le terme composé, formé par inclusion d'un terme plus court, est considéré comme un hyponyme de la tête : *Sauerstofftherapie* est donc un hyponyme de *Therapie*.
La relation d'antonymie est marquée morphologiquement par des préfixes ou des suffixes dérivationnels qui s'opposent sémantiquement. Ainsi, [Schwab *et al.*, 05] décrivent une méthode d'extraction semi-supervisée de couples d'antonymes à partir de leurs préfixes. Les suffixes pourraient également être utilisés mais sont beaucoup plus rares (on peut citer l'opposition *+phile/+phobe*). Le tableau 2 décrit différents types d'oppositions et des exemples de leur réalisation sous forme de préfixes découverts au cours de l'extraction. [Grabar et Hamon, 06][Delphine , 06] proposent également une méthode d'identification de l'antonymie à partir des préfixes de négation comme *dé+*, *ir+*, *anti+*, *non+* ou *in+* et des préfixes privatifs comme *a+* ou *dys+*.
Les liens sémantiques qui s'expriment par des liens morphologiques ne se limitent pas aux seules relations d'inclusion et d'opposition. On trouve d'autres liens sémantiques comme par exemple :
– la répétition : préfixes *re+* ou *ré+* ;
– le changement d'état : suffixes *+iser* ou *+ifier* ;
– la localisation spatiale : préfixes *sur+*, *sous+*, *contre+*, *péri+* ;
– la localisation temporelle : préfixes *pré+*, *post+* ;
– les rôles sémantiques : agent (suffixe *+eur*), résultat (suffixe *+ure*) ;
– la méronymie : suffixes *+age*, *+ade*.
Ces régularités de correspondance entre structure morphologique et liens sémantiques autorisent une approche de l'analyse morphologique de type morphosémantique [Namer, 03].

Type d'opposition	**Préfixes**	**Exemples**
Opposition de degré	hypo+/hyper micro+/macro+ sous+/sur+ infra+/supra+	hypocalorique/hypercalorique microcristaux/macrocristaux sous-alimentation /suralimentation infra-centimétrique/supra-centimétrique
Opposition de nombre	mono+/poly+ uni+/omni+ uni+/bi+ uni+/tri+	monogénique/polygénique unidirectionnel/omnidirectionnel unilingue/bilingue unicolore/tricolore
Opposition dans l'espace	ex+/in+ exo+/endo+ extra+/intra+	exhalation/inhalation exogène/endogène extra-cellulaire/intra-cellulaire
Opposition dans le temps	pré+/post+	pré-caldeira/post-caldeira

Tab. 2: Récapitulatif des relations d'opposition morphologiquement marquées.

3.5.5. Evaluation des résultats d'acquisition de relations sémantiques

Tout comme pour l'extraction terminologique, les résultats des différents systèmes d'acquisition de relations sémantiques sont difficiles à évaluer. Les méthodes d'évaluation sont variées :
– Comparaison à des ressources de référence comme WordNet [Hearst, 1992, Lin, 1998, Widdows et Dorow, 2002, Ferret, 2004], le *Roget's Thesaurus* [Lin, 1998], UMLS [Bodenreider *et al.*, 2001] ou le MeSH [Grabar et Zweigenbaum, 2002a] [Delphine , 06].
– Comparaison aux résultats de tests de vocabulaire à choix multiple, comme les tests du TOEFL (Test of English as a Foreign Language) [Landauer *et al.*, 1998] [Delphine , 06].
– Corrélation aux temps de réaction des sujets pour des tâches de décision lexicale en psycholinguistique [Lund et Burgess, 1996][Delphine , 06].
– Évaluation manuelle [Caraballo, 1999, Buitelaar et Sacaleanu, 2002, Cederberg et Widdows, 2003, Schwab *et al.*, 2005, Wandmacher, 2005][Delphine , 06].

Les deux méthodes d'évaluation principalement utilisées sont la comparaison à des ressources de références et la validation manuelle. La première suppose la disponibilité de telles ressources pour la langue et le domaine traité, tandis que la seconde n'est possible que pour un petit nombre de données.

3.6. Syntagmes et la recherche d'information :

3.6.1. Notion de syntagme :

« Un syntagme est un ensemble de mots formant une seule unité catégorielle et fonctionnelle, il conserve sa signification et sa syntaxe propres» [Khelif, 06].
Il s'agit donc d'un groupe de mots formant une unité à l'intérieur de la phrase. Dans le cadre de l'analyse syntaxique d'une phrase, il s'agit d'une segmentation en unités fonctionnelles appelées syntagmes. Par exemple, on peut citer les types de syntagme suivants : syntagme nominal, syntagme verbal, syntagme adjectival, etc.

Un groupe de mots est souvent plus riche sémantiquement que les mots qui le composent pris séparément. En effet, à titre d'exemple, le terme composé "pomme de terre" est plus précis que "pomme" et "terre" pris isolément. Cet argument a conduit à considérer les groupes de mots comme unité de base dans le langage d'indexation.

Nous avons choisi de nous intéresser aux informations exprimées dans un syntagme défini sommairement comme un ensemble de termes respectant des lois de la morphologie et de la syntaxe, et possédant une signification propre.

Tous les syntagmes partagent un certain nombre de caractéristiques, mais l'essentiel est sans doute le fait que tous ont une tête, c'est-`a-dire, un élément central qui contrôle les autres. Le contrôle exercé par la tête peut se manifester par exemple par l'accord en nombre ou en genre. La tête est donc considérée comme le noyau du syntagme dont dépendent éventuellement d'autres éléments [Haddad, 02].
Certaines études se limitent aux syntagmes nominaux pour aborder le problème de l'indexation automatique.

Selon M. Le Guern, le syntagme nominal est :
" *L'unité minimale de discours qui a la possibilité de signifier un objet (...) {MAISON},le mot du lexique, ne signifie aucune maison que ce soit, alors qu'il suffit que le discours construise le syntagme {UNE MAISON} pour que soit désigné un objet concret. La fermeture du prédicat par le quantificateur {UNE} le transforme en terme.*" [LE GUERN, 89].
« Maison » en tant que mot du lexique, est considéré par l'auteur comme prédicat libre qui ne suppose aucun univers déterminé. Le lexique concerne les mots indépendamment des choses. Le passage du prédicat libre au prédicat lié est une opération qui consiste à placer le mot du lexique dans un univers de discours.

3.6.2. Utilisation des Syntagmes Nominaux en RI :

Parmis les travaux de la recherche d'information qui ont utilisés les syntagmes on trouve :
- SIDHOM Sahbi dans [Sidhom, 02], dans son travail de thèse 'Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'information : de l'écrit vers la gestion des connaissances'.
La contribution de ce travail de thèse s'inscrit au sein d'un domaine multidisciplinaire regroupant le traitement automatique du langage naturel, l'indexation dans un système d'information documentaire et l'organisation des connaissances autour de l'information écrite. Sa particularité consiste en la mise à disposition d'outils pour le traitement automatique de l'information.
L'objectif est de construire *une Plate-forme d'analyse morpho-syntaxique* pour l'indexation automatique et la recherche d'information. Elle est composée d'un noyau d'indexation automatique (*processus d'indexation*) qui utilise le modèle des syntagmes nominaux comme descripteurs de l'information textuelle. Ces syntagmes sont organisés selon une approche Logique Intensionnelle/Extensionnelle (*processus de classification des connaissances*) qui permet d'ordonner les objets d'une classe et de distinguer les classes de connaissances. A la base de cette dernière propriété, il construit son approche pour la recherche d'information (*processus de recherche d'information*).
Cette Plate-forme d'analyse dans sa logique de fonctionnement sera un outil d'investigation orienté vers l'organisation et la gestion des connaissances écrites.
Dans sa recherche, cet aspect sur l'organisation des connaissances a été conduit dans le but de faire émerger les propriétés linguistiques et le traitement du langage dans une pratique expérimentale sur l'indexation automatique documentaire. Il a montré la nécessité de coordonner d'autres sources et stratégies dans l'exploration de ces propriétés. Il s'agit du mode de raisonnement et de la technique d'exploitation des objets du discours spécifiques à la gestion des connaissances (comme étape préalable à la recherche d'information).
Ces deux derniers aspects (mode et technique) intégrés dans le processus de la présentation et de l'organisation du syntagme nominal offrent des scénarii pertinents pour la recherche d'informations.

- Mohamed Hatem HADDAD [Haddad, 02] dans son travail de thèse ' Extraction et Impact des connaissances sur les performances des Systèmes de Recherche d'Information'.
Il a présenté dans cette thèse une approche de traitement de données textuelles qui associe la finesse d'analyse d'une approche linguistique à la capacité d'une approche

statistique d'absorber de gros corpus. En combinant ces deux approches complémentaires, son objectif est d'extraire des connaissances à partir du texte qui peuvent être utiles à un système de recherche d'information.
L'approche statistique se base sur la fouille de données textuelles. La technique de règles d'association permet d'extraire des connaissances relatives aux termes du corpus et qui permet de définir leur contexte d'utilisation en associant un terme à un ensemble de termes par des relations d'association.
L'approche linguistique se base sur les syntagmes nominaux qu'il considère comme les entités textuelles les plus susceptibles de représenter l'information contenue dans le texte. Elle explicite les contraintes linguistiques nécessaires à l'extraction des syntagmes nominaux et explicite les rapports syntagmatiques entre les composantes d'un syntagme nominal. Ces relations syntagmatiques sont exploitées pour la structuration des syntagmes nominaux. La mesure de la quantité d'information permet d'évaluer le pouvoir évocateur de chaque syntagme nominal, de filtrer les syntagmes nominaux et de comparer les syntagmes nominaux entre eux.

- HO Bao-Quoc [Bao-Quoc, 04] dans son travail de thèse 'Vers une indexation structurée basée sur des syntagmes nominaux (impact sur un SRI en vietnamien et la RI multilingue)'.
Dans cette thèse, il a proposé un modèle de recherche d'information à base de syntagmes nominaux structurés. L'objectif global de la thèse est de définir un système de recherche d'information « réellement multilingue ». Dans un premier temps, il a mesuré l'impact des syntagmes nominaux pour des langues comme le vietnamien où les unités linguistiques porteuses de sens sont les termes composés. Ensuite, il a étudié l'adaptation d'un modèle de RI pour prendre en compte un contexte plus large que le terme isolé classiquement utilisé comme terme d'indexation. De ce fait, HO propose de représenter le terme d'indexation avec un peu plus de sémantique. Pour cela, il a structuré les syntagmes nominaux sous la forme de tête et modifieurs. La traduction des requêtes est alors moins approximative que lorsque l'on traduit simplement les mots-clés de la requête, car le syntagme nominal structuré est lui-même plus précis d'après lui.

- Plus récent, Siham Boulaknadel [Boulaknadel, 06] dans son article 'Utilisation des syntagmes nominaux dans un système de recherche d'information en langue arabe'.
Son étude s'intéresse aux connaissances qui peuvent être extraites du contenu textuel des documents en associant la finesse d'analyse d'une approche linguistique à la capacité d'une approche statistique traitant des corpus de grandes tailles. L'approche statistique se base sur la fouille de données textuelles et principalement sur la technique d'analyse sémantique latente tandis que l'approche linguistique se base sur les syntagmes nominaux qu'elle considère comme des entités textuelles plus susceptibles de représenter l'information contenue dans le texte que les termes simples. Et par une expérimentation, sur une collection de documents arabes spécialisés dans le domaine de l'environnement elle a montré l'impact de l'utilisation des syntagmes nominaux sur la précision d'un système de recherche d'information.

3.7. Conclusion :

Nous avons décrit dans cette partie les diverses méthodes d'acquisition automatique de connaissances lexicales, elles se subdivisent en deux tâches principales:
L'identification des mots-clés et termes, qui représentent les concepts d'un domaine, et l'acquisition de relations sémantiques entre termes. Nous avons montré que les approches sont variées, notamment du point de vue des ressources nécessaires : dans certains cas, les connaissances sont extraites à partir de corpus de textes bruts (approches essentiellement statistiques), dans d'autres cas, les corpus sont pré-traités afin de les étiqueter, voire d'effectuer une analyse syntaxique (approches linguistiques). Les deux approches sont souvent combinées pour améliorer les résultats.
Il faut toutefois noter que la majorité des méthodes décrites mettent l'accent d'une part sur l'acquisition de termes polylexicaux et d'autre part sur un mode de structuration des termes reposant sur les connaissances externes (contexte).
Dans ce qui suit, nous présenterons un point de vue général sur la recherche d'information multilingue (SRI multilingue).

4. Recherche d'Information Multilingue (RIM)

4.1. Introduction :

Avec le développement de l'Internet au niveau mondiale, les échanges de documents s'intensifient entre les pays, les cultures et par conséquent, les corpus contiennent de plus en plus de document écrit dans différentes langues. La recherche devient alors multilingue et doit retrouver tous les documents concernés par un besoin d'information.
Dans la vie réelle, un usager qui soumet une requête en français pourrait aussi être intéressé par des documents en anglais, allemand, etc. S'il utilise un système de RI monolingue, il devrait soumettre plusieurs requêtes dans les langues concernées. C'est d'une part lourd à utiliser, et d'autre part impossible pour certains usagers qui ne maîtrisent pas une langue étrangère, mais s'intéresse quand même à obtenir des documents dans cette langue (par exemple, les gens qui font la veille des produits ou des technologies). À la limite, si l'usager ne peut pas comprendre le document dans une langue, il aura toujours la possibilité de faire traduire par un système de traduction automatique ou par un traducteur humain.
Nous présentons dans cette partie ce qui se cache exactement derrière la Recherche d'Information Multilingue (RIM).

4.2. Contexte de la recherche d'information multilingue :

Avant de commencer il faut d'abord posée la question suivante : Qu'est ce qu'une recherche multilingue ?

- Requête multilingue?
- Base multilingue de documents?
- Document multilingue?

4.2.1. Requête multilingue :

Interrogation monolingue de plusieurs bases de documents monolingues

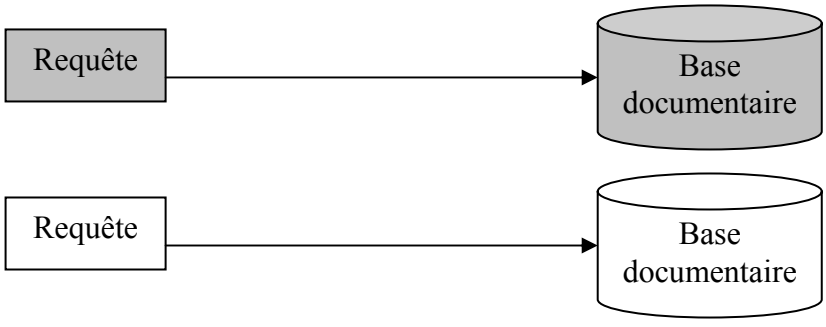


Figure 1.6 : Requête multilingue

Ce genre de système s'apparente à une recherche monolingue car le corpus est découpé en base documentaire monolingue, indépendantes les unes des autres. Les documents de chacune des bases ne peuvent être retrouvés que par une requête dans leur langue.

4.2.2. Base multilingue de documents :

Interrogation multilingue de plusieurs bases de documents monolingues: Système translingue (cross-langage).

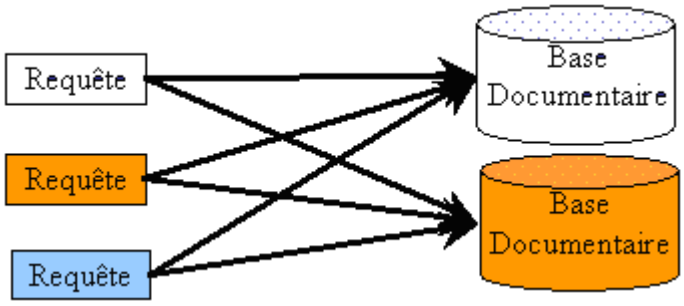


Figure 1.7 : Base multilingue de documents

Donc, à partir d'une requête dans une langue donnée, on peut retrouver des documents écrits dans chacune des langues du corpus.
Ce genre de systèmes Porte le nom « Cross-Language Information Retrieval » (CLIR) appelé aussi système de recherche d'information par croisement de langues.

4.2.3. Document multilingue :

Interrogation de documents multilingues:

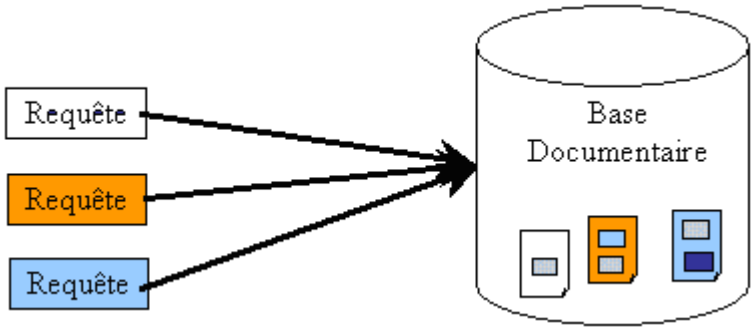


Figure 1.8 : Document multilingue

La recherche s'effectue sur des documents multilingues où des parties du document sont écrites dans des langues différentes. Par exemple, à partir d'une requête dans une langue donnée on peut retrouver des documents multilingues dont le résumé est écrit en français et le corpus de texte en arabe.

En résumé, la recherche d'information multilingue (RIM) est un type de recherche qui permet de repérer l'information lorsque la langue des requêtes est différente de la langue des documents repérés. Un utilisateur peut présenter une requête dans sa propre langue et le système retrouve des documents dans une autre langue. Les principales tâches reliées à la recherche d'information multilingue sont le filtrage, la sélection et le classement des documents qui pourraient être pertinents pour l'usager.
Le principal objectif de la recherche d'information multilingue est de fournir des outils à l'utilisateur qui ne serait pas familier avec une langue particulière, mais qui serait quand même intéressé à obtenir des documents dans une autre langue ou plusieurs autres langues. L'utilisation d'un système de recherche monolingue peut s'avérer fort

problématique pour l'usager lorsqu'il effectue une recherche dans une langue qui ne lui est pas familière. La recherche d'information multilingue tente donc d'apporter une solution à ce problème qui devient de plus en plus préoccupant, depuis l'avènement d'Internet et de son contenu multilingue [Elaine, 04].

4.3. Problèmes de la recherche d'information multilingue :

Comme nous avons cité dans la première partie du chapitre, le processus de recherche d'information à partir d'une requête donnée se compose de deux processus :
Le modèle de représentation des textes (le texte étant à la fois les documents et les requêtes) et le modèle de recherche d'information.
Le premier processus est appelé indexation permettant d'extraire d'un document ou d'une requête, une représentation qui couvre au mieux son contenu sémantique.
Dans le cadre de SRIM, ce processus se complexifie car elle passe obligatoirement par une étape de « traduction » pour représenter le document et la requête dans le même espace d'indexation.
Donc, il est évident qu'une mauvaise traduction des descripteurs entraîne des résultats erronés dans la recherche documentaire. Avant de traduire véritablement un terme, il faut passer par une étape de reconnaissance du concept identifiée par celui-ci. Il ne faut donc plus rechercher des termes mais des concepts [Roussey, 01]. Pour cela, on doit lever au mois trois types d'ambiguïtés sémantiques.
Comme nous avons cité dans la partie de Traitement automatique des langues, parmi les principaux défis de la recherche d'information réside dans les pièges et difficultés du langage naturel.
Dans ce qui suit on récrit ces problèmes :

Polysémie : un même terme peut avoir plusieurs sens.
Par exemple, ''plant'' en anglais possède Trois sens différents en français (la plante, l'installation technique, le coup monté). Donc, si l'on tient compte le contexte des termes, on arrive à déterminer son sens exact.

Homographie : deux mots différents s'écrivent de la même façon.
Par exemple, ''livre'' est soit la conjugaison du verbe livrer, soit le nom synonyme d'ouvrage.

Sens large : un terme qui a un sens très large (*air*) peut prendre un sens particulier dans certains domaines (*air bag*).

Un autre problème de l'indexation multilingue lors de la traduction est traduction concept-concept telle que :

- Structuration différente des concepts.
- Les concepts n'existent pas dans toutes les langues par exemple :
  - Concept de L1 sans équivalent dans L2.
  - Concept de L1 = agrégat de concepts de L2.
  - Concept de L1 ≈ concept de L2.

4.4. Indexation multilingue: approches de traduction automatique :

Il est plus facile de juger de la pertinence d'un document dans une langue étrangère que de formuler une requête efficace dans cette langue. Il est donc plus aisé de pouvoir formuler sa requête dans sa langue maternelle. Un SRI multilingue peut s'appliquer au Web car il est par nature multilingue.
Comme tout SRI, un SRI multilingue doit résoudre le problème de la représentation du contenu des documents dans un index et celui de l'évaluation de la pertinence entre requêtes et documents dans une fonction de correspondance. Cette fonction de correspondance est plus difficile à mettre en œuvre que dans le cas multilingue car la requête et les documents sont dans des langues différentes : on ne peut pas calculer directement la similarité entre eux, comme dans le modèle vectoriel. Il faut passer par une étape de "traduction" des termes, des requêtes ou des documents [Bao-Quoc, 04].
La recherche d'information multilingue est une intersection entre la recherche d'information et la traduction automatique. Le problème de la traduction peut être précisé par les questions suivantes :
1. Quelle est la méthode de traduction utilisée ?
2. Comment choisir une traduction correcte parmi les traductions possibles ?
La première question amène à en poser d'autres : faut-il traduire les documents ou les requêtes ? Quelles sont les ressources utilisées pour la traduction : un logiciel de traduction automatique, un dictionnaire bilingue ou des informations extraites d'un corpus ? Bien que la recherche d'information multilingue hérite des problèmes de la traduction automatique, ils sont plus faciles à régler en RI car on n'a pas besoin d'une traduction exacte pour chaque terme ni d'une traduction lisible et compréhensible par l'utilisateur, donc syntaxiquement correcte. En RI multilingue, la traduction a seulement une contrainte, celle de garder le thème de la requête d'origine. La deuxième question est donc beaucoup moins difficile à résoudre en RI : la notion de « correction » étant beaucoup moins exigeante, et, dans le doute, on peut éventuellement produire plusieurs traductions [Bao-Quoc, 04].
Les approches pour construire un SRI multilingue sont divisées en deux grands types : les approches utilisant un *vocabulaire contrôlé* comme langage d'indexation prédéfini, et celles qui extraient des termes d'indexation du contenu de document (*texte libre*).
Dans les approches basées sur un vocabulaire contrôlé, on utilise une liste prédéfinie de termes d'indexation multilingues pour indexer automatiquement ou manuellement les documents.
La première expérience d'utilisation d'un thésaurus multilingue prédéfini est due à Salton en 1970 [Salton, 70]. Dans son expérimentation, il a utilisé une liste de concepts multilingues anglais – allemand construite manuellement à partir d'une traduction de l'anglais vers l'allemand. L'expérimentation a été réalisée sur un corpus contenant 468 résumés en allemand et 1095 résumés en anglais. La précision moyenne a été d'environ 95% par rapport à celle d'un système monolingue. De cette expérimentation, Salton a déduit que « *cross-language processing…is nearly as effective as processing within a single languag*e » c-a-d «le traitement de croisement de langues est presque aussi efficace que traitant dans une langue simple ».
Il est très rare de disposer, pour indexer les documents, d'un thésaurus multilingue qui couvre plusieurs domaines. Si l'on en dispose, la mise à jour d'un tel thésaurus est coûteuse en temps et délicate pour conserver une cohérence aux termes choisis. C'est pour cette raison que la plupart des travaux en RI multilingue se sont orientés vers l'indexation automatique du contenu de document (texte libre).

Les approches 'texte libre' en RI multilingue sont divisées en trois grands axes. Le premier axe est appelé « *traduction des documents* ». Cette approche utilise un logiciel de traduction automatique ou une traduction manuelle de tous les documents d'une langue vers l'autre. Le deuxième axe propose des méthodes de « *traduction de la requête* ». Ce type d'approche a attiré beaucoup de chercheurs car il est a priori plus facile à réaliser. Le troisième axe utilise un « *langage artificiel* » ou « *langage pivot* » pour représenter les documents et les requêtes qui sont écrits dans des langues différentes [Bao-Quoc, 04].
Ces différenciations découlent du choix de la langue de l'espace d'indexation, c'est-à-dire quelle langue sera employée pour construire les représentations des documents et des requêtes dans le SRI.
d'après les approches pour construire un SRI multilingue peuvent être représentées par la figure suivante :

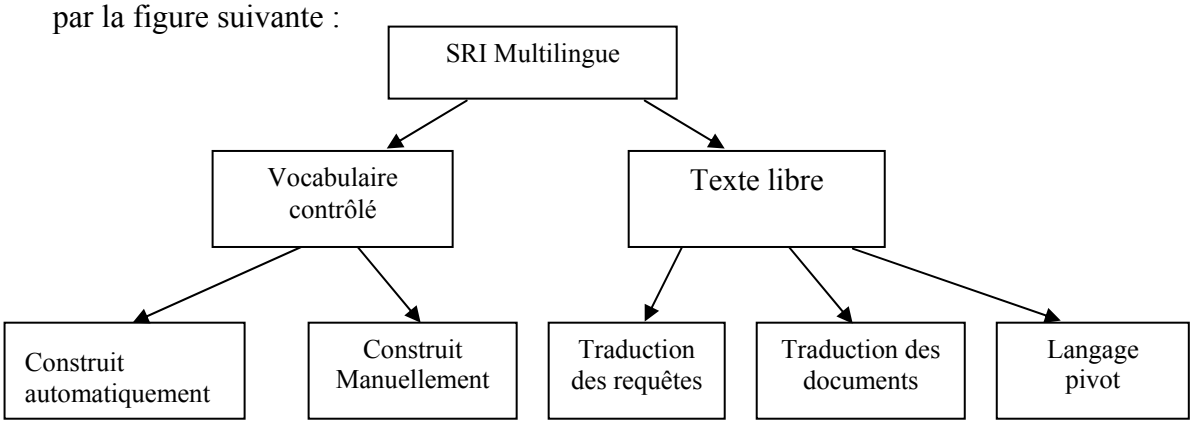


Figure 1.9 : Classification des approches d'un SRI multilingue

Comme cité avant, la plupart des travaux sur la recherche d'information multilingue sont des approches « texte libre ». Dans la suite, on donne une vue générale de cette direction en RI multilingue. Nous présentons dans ce qui suit, les différentes approches connues dans la RIM.

4.4.1. Approche basée sur la traduction de la requête :

Les entités d'indexation appartiennent à la langue du corpus c'est-à-dire les requêtes doivent être traduites dans la langue du corpus avant d'effectuer l'indexation [Roussey, 01].
Cette approche est souvent préférée par les chercheurs en RIM puisqu'il s'agit d'un moyen efficace et peu coûteux, étant donné qu'en général, les requêtes sont composées de mots simples et sont souvent plus courtes et moins complexes à traduire que les documents. Un autre avantage de la traduction des requêtes est qu'il est possible d'intégrer un traducteur de requêtes sans modifier un système de recherche déjà existant.
Mais la difficulté que l'on rencontre pour traduire une requête est le manque de contexte ce qui conduit à interprétations erronées. Cette difficulté réside dans le fait que les requêtes sont généralement composées de quelques mots. Ce manque de contexte est souvent créateur d'ambiguïté ce qui diminue les chances de trouver les bonnes traductions des termes de la requête et par conséquent augmente le bruit généré par le système de recherche.

4.4.2. Approche basée sur la traduction des documents :

Les entités d'indexation appartiennent à la langue de la requête c'est-à-dire le corpus doit être traduit dans la langue des requêtes avant d'effectuer l'indexation [Roussey, 01].
Le principal avantage de cette approche, est d'offrir une plus grande précision de recherche apparente puisqu'un texte plus long pose moins de problème de polysémie lors de la traduction. Donc il permet de compenser le manque de contexte de l'approche précédent.
Cependant, cette technique peut s'avérer longue et forte coûteuse. De plus, s'il faut traduire tous les documents dans toutes les langues, cette tâche apparaît insurmontable et très encombrante en ce qui concerne le stockage (multiplier la taille de la collection par le nombre de langues utilisées par les utilisateurs). Et même, l'alternative d'employer le logiciel de traduction automatique ne fournit pas des résultats suffisants de qualité.
Ces deux approches travaillent en amont du processus d'indexation pour modifier l'une des entrées du processus d'indexation, les documents ou les requêtes. Par conséquent, l'indexation est toujours monolingue même si plusieurs langues sont utilisées pour les requêtes.

4.4.3. Approche basée sur le langage pivot :

Concernant les limites des méthodes précédentes, quelques autres approches proposent d'adresser l'indexation multilingue en employant un langage pivot.
Les entités d'indexation appartiennent à un langage formel c'est-à-dire les documents et les requêtes doivent être traduits dans un langage formel qui n'est pas dépendant d'une seule langue [Roussey, 01].
Cette approche est basée sur une indexation multilingue car les documents et les requêtes sont représentés par des entités d'indexations qui ne sont pas dépendantes d'une seule langue. Le langage pivot entre la langue des documents et celles des requêtes.
Cette approche implique donc une '' traduction '' des documents et de la requête dans ce langage formel. La première remarque est évidente : l'analyse linguistique est donc deux fois plus lourde car il faut ''traduire'' le corpus et les requêtes. Par contre, le passage à une recherche d'information vraiment multilingue, c'est-à-dire travaillant sur une collection de document incluant au moins deux langues est facilité. En effet, dans le cas d'un système CLIR capable d'interroger une collection monolingue de documents dans n langues, différente de celle du corpus, il faut autant de ressources linguistiques que de langues d'interrogation pour les deux premières approches, c'est à dire n. Pour une indexation en langage pivot, il faut n+1 ressources linguistiques, n pour traduire en langage pivot les n langues d'interrogation plus une pour traduire dans la langue de la collection. Par contre si l'on se place dans le cadre d'un système multilingue capable d'interroger une collection des documents contenant n langues avec n même langues, alors les deux premières approches nécessitent l'utilisation de n(n-1) ressources linguistiques pour effectuer les traductions des requêtes ou des documents. Alors qu'avec un langage pivot, nous avons besoin de n ressources permettant le passage de chaque langue vers le langage pivot, comme le montre la figure suivante.

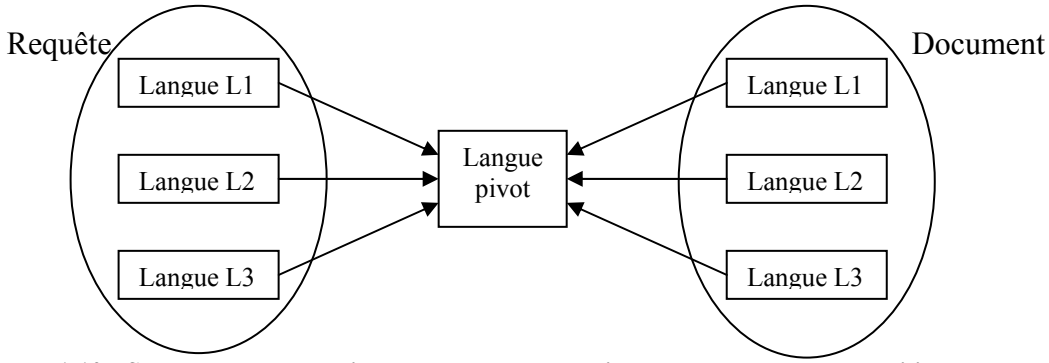


Figure 1.10 : Schéma des traductions avec un langage pivot dans un système multilingue

4.5. Ressources linguistiques pour le traitement d'information multilingue :

Pour représenter les documents et les requêtes dans la langue de l'espace d'indexation, les systèmes multilingues disposent de différentes ressources pour trouver les traductions. Le choix de la ressource linguistique est très important car la qualité des résultas du système est fortement dépendante de la qualité de la ressource utilisée.
Dans ce qui suit, nous les représentons brièvement en donnant leurs caractéristiques :

4.5.1. Système de traduction automatique :

L'utilisation d'un logiciel de traduction automatique est l'approche la plus directe.
Ces systèmes sont utilisés pour obtenir différentes versions d'un même texte dans plusieurs langages or le but d'un système de traduction automatique est de produire une version lisible et fiable dans la langue cible du texte source donc il semble être adapté pour caser la barrière langagière dans un SRIM.
Les problèmes de cette approche sont [Bao-Quoc, 04] et [Roussey, 01]:

1- Choix incorrect de la traduction du mot ou terme :

Un traducteur automatique doit forcément lever touts las ambiguïtés pour ne fournir qu'une unique version de la traduction, ce qui génère parfois des choix incorrects. Malheureusement, les SRI sont plus pénalisés par un mauvais choix de traduction que par la persistance d'une ambiguïté.
Par exemple : « organic food » est traduite par « nourriture organique » alors que la bonne traduction est « nourriture biologique ».

2- Syntaxe incorrecte :

Un traducteur automatique doit fournir en sortie un texte grammaticalement correct. Or les SRI sont plus sensibles à des traductions sémantiquement correctes qu'à des constructions syntaxiquement valides. Un traducteur remplit donc des tâches pas nécessairement utiles pour le SRI.
Par exemple : « human-assisted machine translation » est traduite par « traduction automatique humain-aidée » alors qu'une bonne traduction possible peut être « traduction automatique assistée manuellement ».

3- Traduction des mots inconnus :

La ressource linguistique doit être adaptée aux vocabulaires du corpus et des requêtes ou des documents alors leurs termes ne seront pas reconnus dans l'étape de traduction et ils ne seront pas pris en compte, ni par le processus d'indexation pour représenter un document, ni par la fonction de comparaison pour comparer les représentations. De plus si la ressource linguistique ne permet pas de trouver les traductions correctes des expressions alors ces erreurs seront répercutées sur la comparaison et le système de recherche proposera des résultats erronés.
Par exemple : le nom propre Bérégovoy peut être traduit par Bérégovoy ou eregovoy.
Les noms propres en particulier et les entités nommées en général doivent subir un traitement particulier. Il faut qu'ils soient identifiés surtout s'ils ont un sens possible comme "Bill Gates" ("Les portes de la facture") ou s'ils doivent trouver un équivalent phonétique dans la langue cible comme c'est le cas par exemple pour le chinois vers l'anglais.
Voici quelques systèmes de traduction automatique qui ont fait leurs preuves d'après [[Elaine, 04] :

- Systran :
Systran est un fournisseur de service de traduction automatique facilitant la communication pour 36 paires de langue et dans 20 domaines spécialisés. Systran est utilisé pour de nombreuses applications : commerce électronique, gestion de bases de données, intranets d'entreprise, services de messagerie, etc. Plusieurs sociétés (Ford Motor Company, Cisco Systems, Daimler Chrysler Corporation, Price Waterhouse Coopers, etc.) et portails (Google, AOL, Altavista, Lycos, Oracle, etc.) utilisent l'expertise de Systran.

- CAT2 :
CAT2 est un système de traduction automatique multilingue qui a pris ses racines dans le projet EUROTRA. CAT2 offre des caractéristiques telles que la possibilité d'obtenir une traduction robuste et la possibilité d'un codage rapide par le traitement de lacunes du lexique (dictionnaire). Ce système de traduction automatique a été développé afin de pouvoir s'adapter aux besoins de traduction de nombreuses entreprises. De plus, CAT2 est utilisé pour l'enseignement de l'usage de dictionnaires, de règles de syntaxe et de transfert aux futurs traducteurs dans plusieurs universités européennes.

- Reverso :
Reverso est un système de traduction automatique de la compagnie Softissimo. Ce système permet de traduire des documents peu importe leur taille ou leur format (Excel, Word, etc.), tout en respectant la mise en page. Reverso est basé sur une technologie rapide et efficace et a été adopté par de nombreuses entreprises (France Telecom, Renault, Nestlé, etc.). En plus d'une interface conviviale, Reverso propose de nombreux outils de personnalisation linguistique, de révision et de prononciation. Reverso est offert dans de nombreuses versions pour les particuliers et les professionnels. De plus, ce système de traduction automatique peut être enrichi par une gamme de dictionnaires spécialisés pour le vocabulaire très technique.

- Logomedia :
Logomedia offre un système de traduction automatique des documents, sites Web, courriers électroniques, etc. La traduction s'effectue de et vers l'anglais et les principales langues européennes et asiatiques. De plus, les utilisateurs peuvent combiner de nombreuses paires de langues. Ce système offre aux usagers un accès facile à une base de données linguistiques importante et à un logiciel sans surcharger leur ordinateur. Logomedia propose d'étendre leurs services aux sessions de clavardage, aux messages courts destinés au WAP (téléphone cellulaire) et aux documents fax.

4.5.2. Les bases lexicales :

Différents types de base lexicale sont utilisés pour trouver les traductions d'un terme. Le contenu d'une base lexicale peut aller de la liste des traductions d'un terme en entrée, aux contraintes de sélection des traductions. Plus la base contient des informations pertinentes sur l'utilisation des traductions, plus les traductions trouvées seront de qualité. Nous ne présentons que deux types de base lexicale : la plus simple, le dictionnaire de transfert et la plus complète, les ontologies (bases de connaissances).

4.5.2.1. Les dictionnaires de transfert :

Un dictionnaire de transfert fournit en sortie les traductions d'un terme donné en entrée.
Les dictionnaires bilingues de transfert sont les données les plus couramment accessibles par le biais des dictionnaires électroniques. Par contre, ceux-ci ne peuvent pas être utilisés tels quels car ils ont besoin de subir certaines transformations de leurs données afin de les rendre plus adéquates aux besoins de la recherche d'information [Huull, 96][Rroussey, 01]. En effet, les dictionnaires sont destinés à un utilisateur humain et non à un système informatique ; d'où les quelques problèmes générés par l'utilisation des dictionnaires électroniques.
Une étude sur des approches se basant sur un dictionnaire bilingue a été réalisée par Pirkola et al. [Pirkola, 01][Bao-Quoc, 04]. Les problèmes principaux associés à cette approche sont :
1- la couverture de dictionnaire : c'est à dire des termes ne peuvent pas être traduits parce qu'ils n'existent pas dans le dictionnaire ou autrement dit, comment traiter des termes non traduit parce que ce sont des termes composés nouveaux, des noms propres, etc.
Un dictionnaire ne contient pas tous les mots possibles que nous puisions trouver dans un texte. Certaines formes d'un terme sont explicites pour un lecteur et ne seront donc pas nécessaires dans un dictionnaire, car un utilisateur humain sera capable de dériver automatiquement ces formes : par exemple, trouver l'adverbe à partir de l'adjectif.
Les dictionnaires couvrent un vaste domaine de connaissances et donc certains termes spécifiques à un domaine particulier n'ont pas leur place dans un dictionnaire générique.
Les dictionnaires contiennent des définitions longues, avec beaucoup de bruit. Par exemple, le verbe « prendre » a 23 synonymes en anglais. Toutes ses traductions possibles n'ont pas forcement le même sens (cas des termes polysémiques), donc en tenant compte de chaque traduction possible, on augmente le bruit de la recherche documentaire.

Les dictionnaires contiennent des définitions encyclopédiques, contenant des mots contextuels, comme « quelqu'un » ou « chaque chose », inadaptés à la recherche documentaire.
2- l'ambiguïté de la traduction : il faut choisir la traduction la plus correcte dans le contexte où est utilisé le terme.
Par conséquent, ce genre de dictionnaire doit être nettoyé manuellement avant d'être utilisé.
Il existe environ 6800 langues parlées dans les 191 pays du monde. Parmi ces langues répertoriées, 2261 possèdent un système écrit. Un dictionnaire en ligne existe pour environ 260 de ces langues. Quelques exemples de ces dictionnaires sont décrits ici d'après [Elaine, 04] :

- YourDictionary :
YourDictionary.com est une entreprise de produits et services dont le portail fournit plus de 2500 dictionnaires et grammaires pour plus de 300 langues. Parmi les dictionnaires offerts par YourDictionary, on retrouve des dictionnaires multilingues généraux, de même que de nombreux dictionnaires terminologiques sur des sujets variés, de la biologie à la cuisine en passant par la religion. Ce site semble des plus complets et assez convivial pour être utilisé aussi bien par le profane que par le spécialiste.

- Freelang :
Le dictionnaire bilingue Freelang est un dictionnaire distribué gratuitement sur le site Web de Freelang. Ce dictionnaire utilise le français comme langue principale. Il comporte une liste de mots se composant de deux colonnes : une liste de mots d'une langue étrangère, et une liste de traductions correspondantes dans la langue principale. Une centaine de langues sont représentées et il est intéressant d'y retrouver certaines langues comme l'alsacien, le niçois et l'arpitan savoyard. Le principal avantage de ce site est sans aucun doute son interface en langue française.

- Ultralingua :
Ultralingua est une entreprise qui développe des logiciels linguistiques dans le domaine des affaires et de l'éducation. Ultralingua offre une variété de dictionnaires en ligne pour de nombreuses paires de langues, mais propose également la possibilité de conjuguer des verbes, d'écrire des nombres en plusieurs langues et de consulter des grammaires regroupant les principales difficultés d'une langue spécifique. Ce site nous semble intéressant mais tout de même limité dans le choix de langues offertes.

- Babylon :
Babylon offre 25 dictionnaires complets exclusifs, en 13 langues, et contenant plus de 3 millions de mots et de phrases, de même que plus de 1600 glossaires en 70 langues et sur des sujets variés comme les sciences, la culture et les sports. De plus, Babylon offre des fonctions de traduction bidirectionnelles permettant à l'usager de traduire de l'anglais vers 13 langues et vice-versa.

- LangTolang :
LangToLang est un service Internet qui offre un accès à plusieurs dictionnaires multilingues en ligne gratuitement. Ce service propose également plusieurs fonctionnalités intéressantes. Parmi celle-ci, mentionnons la possibilité de télécharger

des dictionnaires multilingues pour les téléphones portables. Ce service, évidemment, n'est pas gratuit!

4.5.2.2. Utilisation des bases de connaissances : ontologies et thésaurus :

Une ontologie est un ensemble de termes formels, auxquels peuvent être associées des définitions, permettant de représenter des connaissances. Ces définitions permettent :

- De poser des contraintes sur l'utilisation des termes et ainsi d'effectuer des vérifications syntaxiques ou sémantiques, autrement dit de préciser le contexte d'utilisation du terme.
- De guider l'association de certains termes avec d'autres et d'indiquer également des relations possibles entre termes.

Le thesaurus peut être considéré comme une forme particulière d'ontologie. Dans un thesaurus multilingue, la relation d'équivalence inclue dans l'ensemble des termes choisis comme représentant du concept, leurs traductions et leurs synonymes. La relation d'association va nous permettre d'améliorer la traduction des expressions. En effet, d'abord chercher une traduction mot en considérant les termes comme indépendants, il faut d'abord chercher à les relier par des relations d'associations pour obtenir la traduction exacte d'un concept multiterme [Roussey, 01].

Un thésaurus multilingue utilisant un vocabulaire contrôlé pour l'indexage et la recherche peut être envisagée comme un ensemble de thésaurus monolingues qui englobent tous un même système de concept. Dans un vocabulaire contrôlé, on a un ensemble défini de concepts utilisés pour l'indexage et la recherche. De cette manière, le problème de l'ambiguïté est éliminé. Les utilisateurs peuvent employer un terme dans leur langue maternelle pour établir le concept identifiant correspondant afin de retrouver les documents dans d'autres langues. Dans les systèmes les plus simples, ceci peut se réaliser manuellement en consultant un thésaurus qui inclut pour chaque concept les termes correspondants en plusieurs langues et qui possède un index par langue. Dans les systèmes plus élaborés, la relation entre terme et descripteur sera réalisée en interne [Carol, 01].
Dans l'approche fondée sur un vocabulaire contrôlé, les termes appropriés du vocabulaire doivent être assignés à chaque document de la collection. Traditionnellement, ceci était fait manuellement par des experts du domaine. C'est très coûteux. On développe actuellement des méthodes semi-automatiques pour placer ces indicateurs. Il reste que les thésaurus sont très chers à construire, coûteux à maintenir et difficiles à mettre à jour. De plus, il n'est pas facile de former les utilisateurs pour qu'ils utilisent correctement les relations du thésaurus.
Donc le problème des ontologies et thesaurus réside dans leur construction laborieuse, leur maintenance et leur mise à jour onéreuses. De plus, il est difficile d'établir exactement des équivalences entre des concepts de langues différentes, surtout lorsque plus de trois langues sont en jeu.

4.5.3. Les corpus :

L'approche fondée sur les corpus analyse de grandes collections de textes sur une base statistique et extrait automatiquement l'information nécessaire à la construction de techniques de traduction propres à une application. Les collections analysées peuvent être constituées de textes parallèles (traduction équivalente) ou comparables (lié à un

domaine). Les démarches principales utilisant les corpus sont l'espace vectoriel et les techniques probabilistes.

4.5.3.1. Les corpus parallèles :

La technique de traduction utilisant un corpus de textes parallèles consiste à traduire les requêtes à l'aide de termes extraits de collections de documents parallèles, c'est-à-dire le même texte traduit en une ou plusieurs langues. Cette technique suppose que la recherche d'information est utilisée pour trouver un sous-ensemble de textes correspondant à une requête particulière. Les textes parallèles sont également utilisés pour produire une traduction. Le corpus parallèle est utilisé afin de désambiguïser les termes équivalents d'une langue (langue cible) en comparant les résultats de recherche aux termes de la requête effectuée dans une autre langue (langue source) [Carol, 01].
Donc un corpus parallèle est une base de documents contenant pour chacun d'eux leur traduction dans chacune des langues du corpus.

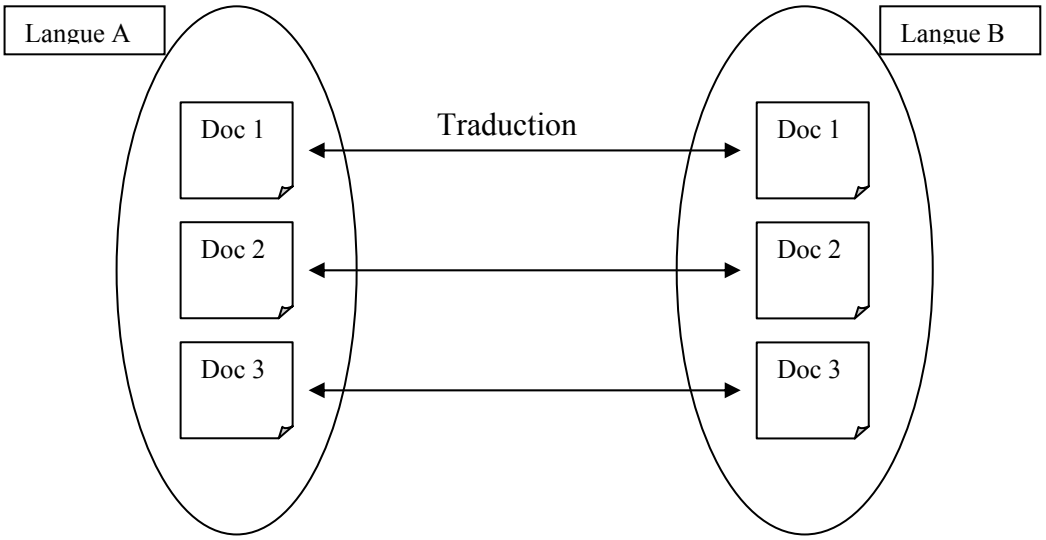


Figure 1.11 : Schéma d'un corpus parallèle.

4.5.3.2. Les corpus comparables :

Une collection de documents comparables est constituée de documents rassemblés sur la base de la similarité des sujets traités plus que sur leur équivalence en traduction. L'idée qui sous-tend l'utilisation de ces corpus est que les mots utilisés pour décrire un sujet particulier seront liés sémantiquement à travers les langues [Carol, 01].
Donc un corpus comparable est un ensemble de documents traitant d'un même domaine. Un document n'a pas forcément sa traduction exacte dans le corpus [Roussey, 01].

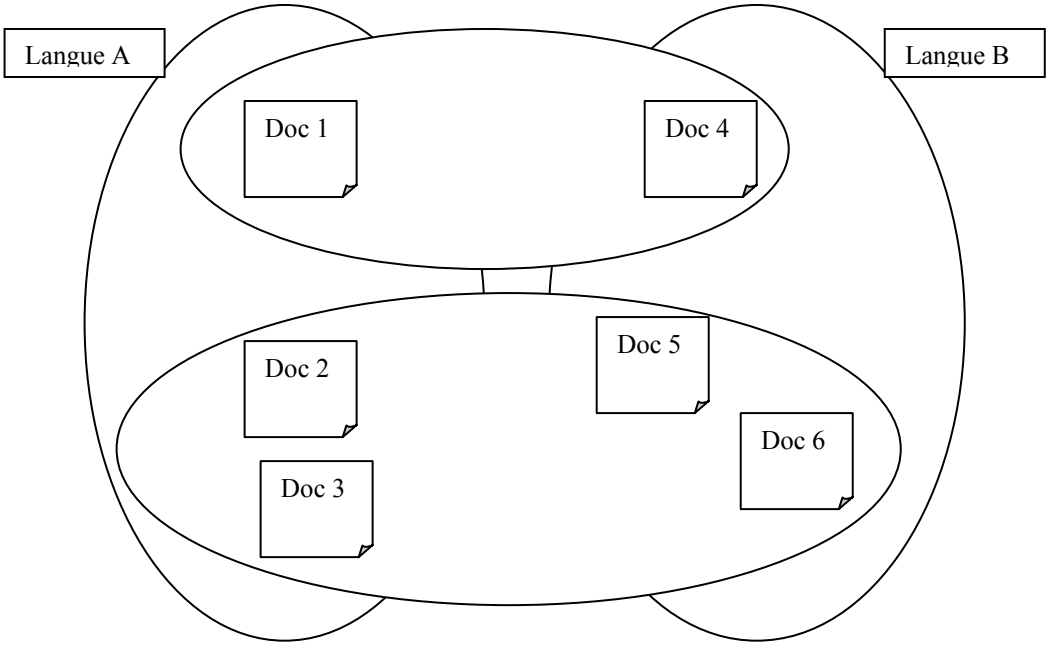


Figure 1.12 : Schéma d'un corpus comparable.

Finalement, les corpus parallèles sont des ensembles de textes traduits équivalents, constitués généralement du texte source et d'une ou plusieurs traductions, les corpus comparables sont plutôt des textes pouvant être associés, non par leur traduction mais plutôt par leurs caractéristiques communes (même sujet, par exemple).
L'utilisation de corpus parallèles et comparables est une alternative qui a soulevé beaucoup d'intérêt chez les chercheurs en RIML. Plusieurs corpus ont ainsi été constitués. En voici quelques exemples d'après [Elaine, 04]:

- Multext :
Multext est un projet visant le développement d'outils, de corpus et autres ressources pour une vaste variété de langues incluant le bambara, le bulgare, le catalan, le tchèque, le hollandais, l'anglais, le français, etc. Le corpus de Multext est développé sur une base volontaire et est disponible au grand public à des fins non commerciales ni militaires. Deux corpus ont été utilisés pour le développement des outils linguistiques de Multext : un ensemble de documents tirés du Official Journal of European Community en cinq langues et un corpus comparable de textes de nouvelles financières en suédois.

- MultiTrans :
MultiTrans est un corpus multilingue plein texte intégré à une infrastructure évoluée de gestion terminologique. MultiTrans se veut un outil de soutien à la traduction. Il permet de transformer rapidement des traductions préalables et d'autres contenus en banques d'expressions de toute longueur pouvant être consultées dans leur plein contexte. L'alignement et l'indexaation plein texte se font automatiquement, de même que l'extraction des expressions répétitives. Une fonction de recherche permet également l'identification rapide de toutes les occurrences d'une expression, la consultation dans leur contexte initial et l'affichage des traductions préalables.

- European Corpus Initiative Multilingual Corpus :
Le corpus ECI (European Corpus Initiative Multilingual) est un corpus contenant plus de 98 millions de mots assurant la couverture de la majorité des langues européennes, de même que la langue turque, le japonais, le chinois, le malais et plusieurs autres. Le principal objectif de ce corpus est la cueillette de données textuelles de toute sorte, incluant la transcription de matériel parlé. Le corpus ECI existe en version CD-ROM depuis 1994 et est distribué par ELSNET. Ce corpus regroupe des textes tirés de journaux et de documents scientifiques.

4.6. Exemple d'un SRIM:

SyDoM : Système Documentaire Multilingue (Roussey Catherine):

C'est un outil d'annotation pour le web sémantique. Est un système répondant à différents besoins du web et permet l'amélioration de la représentation du contenu des pages web et la recherche multilingue. Il est adapté à la gestion de documents textuels stockés aux formats XML. SysDoM est composé de trois modules :

4.6.1. Le module de gestion du thésaurus sémantique :

Un thésaurus sémantique est un nouveau genre d'ontologie. Ce module permettant de construire un langage documentaire utilisé pour annoter et interroger les documents XML. Ce langage se compose d'une modélisation du domaine à laquelle sont associés plusieurs vocabulaires. Un thésaurus sémantique définit deux niveaux de connaissances :

4.6.1.1. Le niveau conceptuel (support):

modélise le domaine d'étude formé de types de concepts ou de relations. Ce niveau ne dépend pas d'une seule langue (type ≠ terme). Dans ce niveau il y a la définition du langage pivot (indexation en langage pivot).

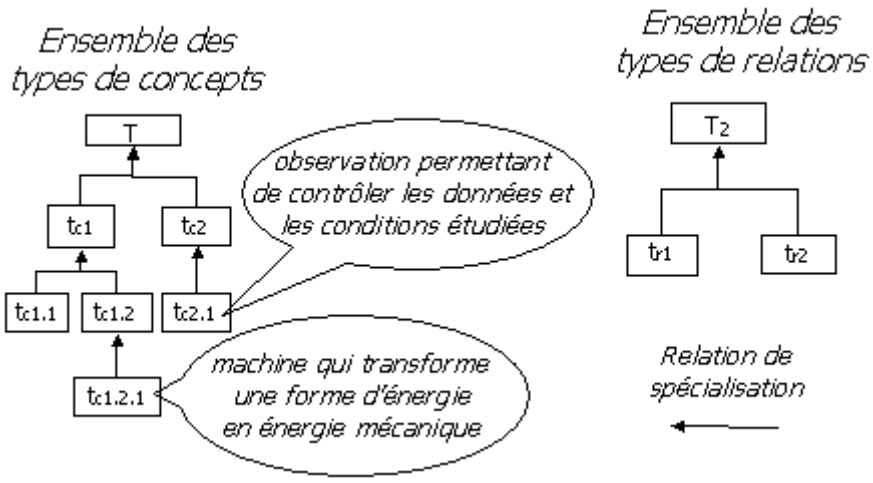


Figure 1.13 : Thesaurus Sémantique : Niveau conceptuel

4.6.1.2. Le niveau terminologique :

Un vocabulaire est l'ensemble de termes d'une langue. Ce niveau est composé de l'ensemble des termes, le terme étant défini comme la manifestation linguistique d'un concept repéré dans un texte c'est à dire le terme dans un contexte référence un concept (terme=label d'un type). Il définit les langages de représentation pour l'utilisateur.
Comme le thésaurus sémantique est l'élément central du système SyDoM, la création de ce thésaurus est une étape importante pour les processus d'indexation et de recherche. Ce module permet de construire une modélisation du domaine en créant une hiérarchie de types de concepts et de types de relations. La création des hiérarchies de types est accompagnée de la création des terminologies permettant de lier un ensemble de termes à un type.
Ce module de SyDoM permet de renseigner l'ensemble des informations nécessaires à la création d'un type de concepts ou de relations. Un type se définit par :
Un identifiant numérique caractérisant le type de manière unique. Par exemple, '$t_{c1.1.1.1.1.1.5}$' est l'identifiant d'un type de concepts.
Des définitions en langage naturel. Par exemple, le type de concepts '$t_{c1.1.1.1.1.1.5}$' représentant la notion de moteur à réaction a pour définition en français "moteur produisant un courant rapide de gaz chaud qui propulse le véhicule qu'il équipe". Le thésaurus étant multilingue, il est nécessaire de créer une définition pour chaque langue présente dans le thesaurus sémantique.
Des listes de termes appartenant aux différentes terminologies du thesaurus sémantique sont associées aux types. Par exemple, le terme "réacteur" appartient à la liste des termes français associée au type '$t_{c1.1.1.1.1.1.5}$' identifiant la notion de moteur à réaction. Pour caractériser le lien entre le terme et le type, un poids est associé à chaque terme. Ainsi un terme de poids fort sera jugé plus représentatif du type, qu'un terme de poids faible.
Dans le cas des relations, la signature d'un type de relations spécifie le type de concepts le plus générique que peut avoir un argument d'une relation de ce type. Par exemple, une relation de type ' $t_{r1.1}$' a pour signature $\sigma(t_{r1.1})=(t_{c2}, t_{c1})$. Par conséquent, une relation de ce type doit avoir comme premier argument un nœud concept dont le type spécialise 't_{c2}'.
Un type de concepts ou de relations se définit par les relations de spécialisation et de généralisation qu'il entretient avec les autres types dans une hiérarchie. C'est-à-dire qu'un type peut généraliser et spécialiser plusieurs autres types. Dans sa hiérarchie, ce type aura donc plusieurs parents et plusieurs enfants. Dans le cas des types de relations, la position de la hiérarchie inclut la vérification de la cohérence des signatures, c'est-à-dire que les types de concepts, appartenant à la signature d'un type de relations, doivent être plus spécifiques que les types de concepts de la signature du type parent.
Ce module permet non seulement de construire une ontologie, mais aussi de modifier les thésaurus sémantiques existants, en insérant ou supprimant de nouveaux types ou en associant de nouveaux termes aux types existants. Dans le cas des types de concepts, la suppression d'un type n'est possible que si ce type n'apparaît pas dans la signature d'un type de relations. De plus, toute modification d'un type n'est autorisée qu'à condition que cette modification n'engendre pas d'incohérences dans la base des graphes indexant un document.

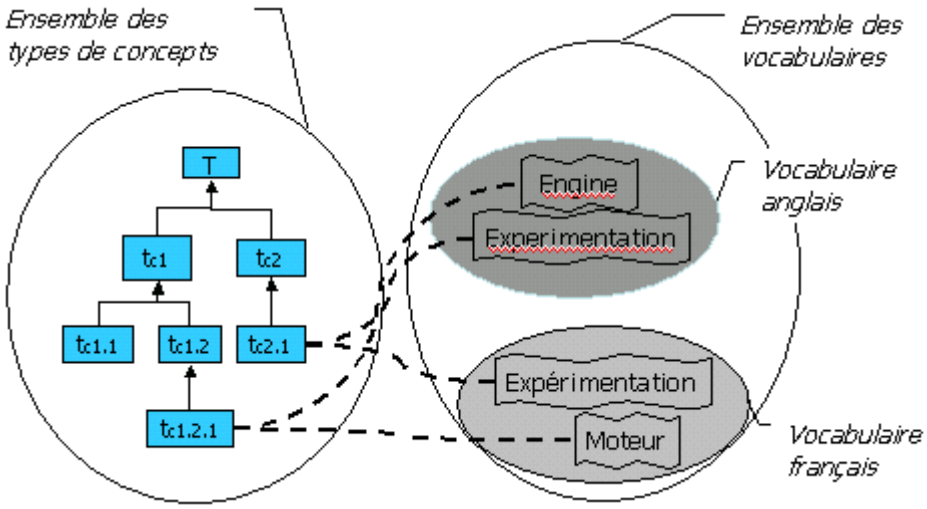


Figure 1.14 : Thesaurus Sémantique : Niveau terminologique

Le thésaurus sémantique doit être présenté aux différents utilisateurs de SyDoM, dans la langue de leur choix. Par conséquent, un navigateur permet de parcourir les hiérarchies de types du thésaurus, en sélectionnant une langue d'affichage, où seules les informations écrites dans cette langue sont présentées. De ce fait, l'utilisateur visualise les types au moyen des termes et des définitions d'une langue donnée associés aux types. La figure 1.15 présente deux versions du même thésaurus sémantique. La version anglaise présente tous les types de concepts et de relations à l'aide de termes anglais. La version française du thésaurus sémantique propose les mêmes hiérarchies de type avec des termes français. Les types étant présentés dans la hiérarchie par leur ordre alphanumérique, tout changement de langue peut avoir des répercutions sur l'ordre des types appartenant aux même niveaux hiérarchiques.

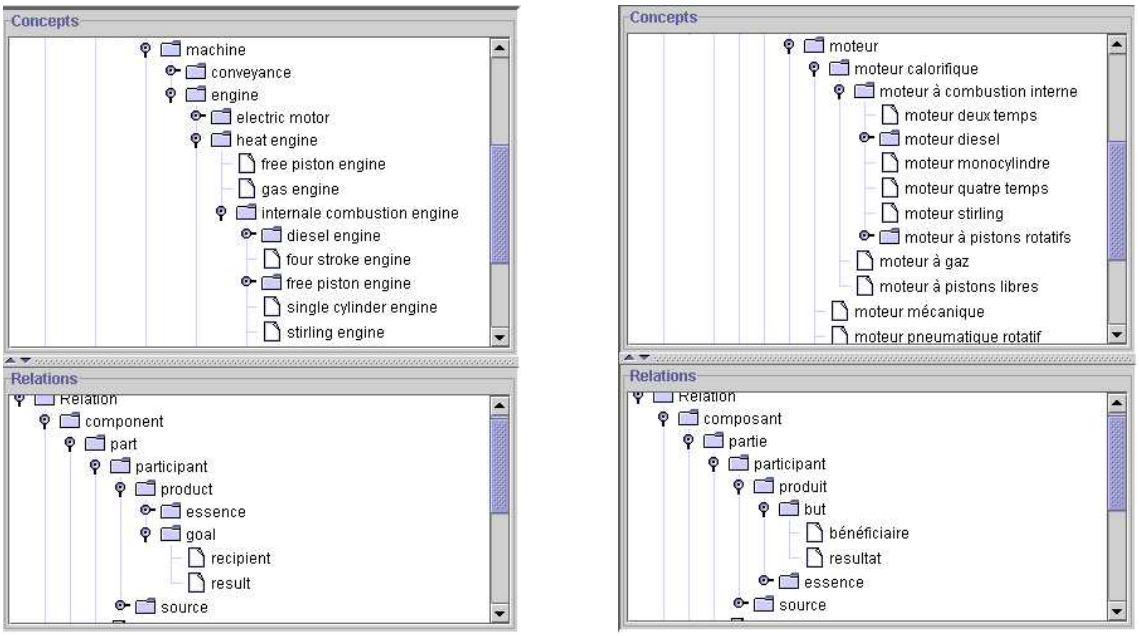


Figure 1.15 : La version anglaise et française d'un thésaurus sémantique.

4.6.2. Le module d'indexation :

Ce module permet d'annoter les documents par des graphes conceptuels.
Ce module de SyDoM a pour but d'enrichir une page web pour faciliter son utilisation ultérieure. Cet enrichissement consiste à insérer des connaissances, issues de l'interprétation de la page web, dans son contenu. Le format de nos pages étant XML, nous avons défini une série de balises sémantiques. Ces balises nous permettent d'associer des connaissances à des parties de document, jugées pertinentes par l'utilisateur. De plus, les annotations permettent de compléter le thésaurus sémantique pour améliorer l'adéquation entre les terminologies du thésaurus et celles des pages web [Roussey, 01].
Le processus d'indexation des pages web tient compte des deux types de connaissances définies dans le thésaurus sémantique :

4.6.2.1. Les annotations :

Annoter c'est analyser et interpréter le contenu du document c'est à dire identifier les termes importants, identifier l'occurrence de terme dans son contexte, interpréter le sens de l'occurrence dans son contexte et associer un graphe composé d'un seul nœud concept à l'occurrence (associer une occurrence d'un terme à un concept).
Les annotations représentent une indexation à partir des terminologies. Ces annotations identifient un terme dans son contexte, comme représentant d'un concept particulier. Les concepts référencés par des termes constituent un graphe conceptuel. Ces annotations permettent d'enrichir automatiquement les terminologies du thésaurus sémantique.

4.6.2.2. L'index du document :

Indexer c'est sélectionner les informations importantes c'est à dire déduire des annotations les concepts les plus importants, lier ces concepts entre eux par des relations et construire un graphe sémantique, et enfin, représenter ces graphes en langage pivot pour ne pas dépendre uniquement de la langue du document.
L'index du document est une indexation à partir des connaissances du domaine. L'index est un raffinement du graphe issu des annotations : Il est composé de nœuds concepts associés entre eux par des nœuds relations. L'index est non seulement une mise en relation d'instances de concepts utilisés dans les annotations, mais ces instances sont le résultat d'un processus de filtrage des annotations, car cet index se veut une synthèse du contenu du document. Les index pourront être associés au thésaurus sémantique pour aider la compréhension des types.

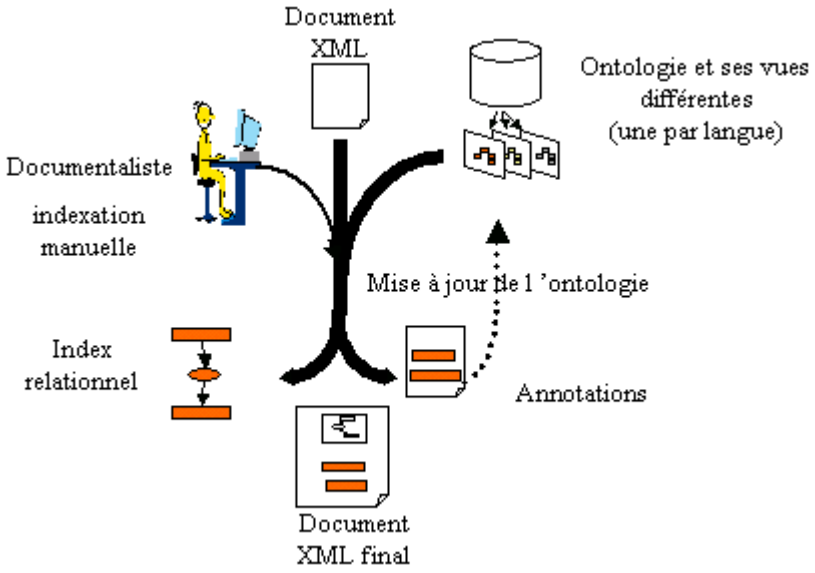


Figure 1.16 : Système d'indexation

Comme le montre la figure 1.17, ce module se compose, sur la partie droite, d'un navigateur permettant de parcourir le thésaurus sémantique, et sur la partie gauche d'un éditeur de graphes conceptuels. Pour construire un graphe indexant un document XML, l'utilisateur parcourt le thésaurus sémantique dans la langue de son choix, et sélectionne des types de concepts ou les types de relations. Une fois le graphe conceptuel construit, le document XML est automatiquement enrichi en insérant une série de balises sémantiques au début du document. Ces balises représentent le graphe dont chaque nœud concept et chaque nœud relation est identifié par l'identifiant de son type. D'autres balises sémantiques caractérisent les termes dans leur contexte en leur associant un nœud concept. Ensuite, la base de données contenant les index et les annotations est mise à jour.

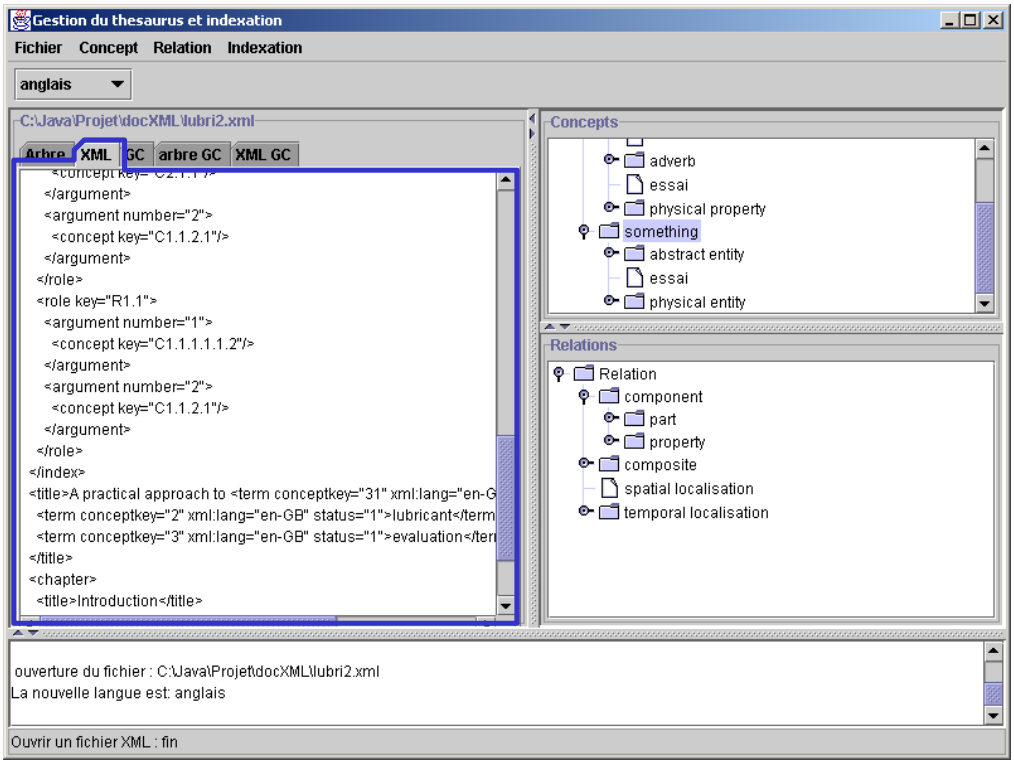


Figure 1.17: L'interface du module d'annotation de SyDoM.

4.6.3. Le module de recherche :

Le module de recherche contient les mêmes composantes que le module d'indexation, c'est-à-dire un éditeur de graphes et un navigateur pour parcourir le thésaurus sémantique. L'éditeur de graphes permet à l'utilisateur de construire une requête dans la langue de son choix en sélectionnant des types, dans le thésaurus sémantique.

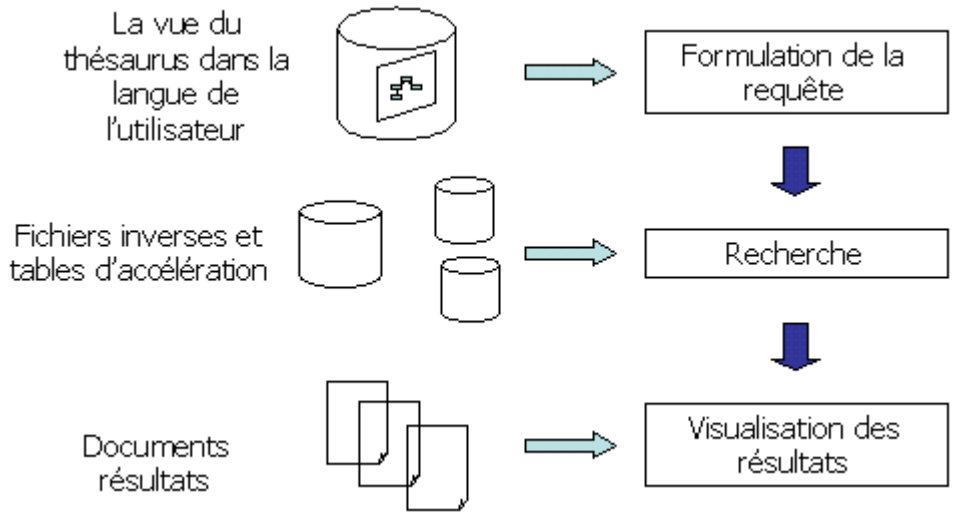


Figure 1.18 : système documentaire multilingue

Par exemple, la figure 1.19 présente un graphe requête correspondant au besoin d'information sur "la génération de bruit dans les moteurs diesels". Une fois le graphe requête construit, celui-ci est automatiquement transformé par le système pour que les nœuds des graphes soient étiquetés par les identifiants numériques de leur type et non par des termes. Dans le prototype SyDoM, la fonction de comparaison entre les graphes index et les graphes requêtes compare les arcs des graphes, c'est-à-dire un nœud relation liant plusieurs nœuds concepts. Autrement dit, les graphes sont décomposés en un ensemble d'arcs. Le poids des documents résultat de la figure 1.19 correspond à la similarité entre arcs du graphe requête et des graphes index. Les fonctions de similarités utilisées sont présentées en détail dans [Roussey, 01].En résumé, ces fonctions de comparaison évaluent une similarité entre les documents et les requêtes, ce qui permet de constituer, pour chaque requête, la liste des documents jugés les plus pertinents. Cette liste de documents est ensuite affichée à l'utilisateur, chaque document étant pondéré par la similarité entre son graphe index et le graphe requête.

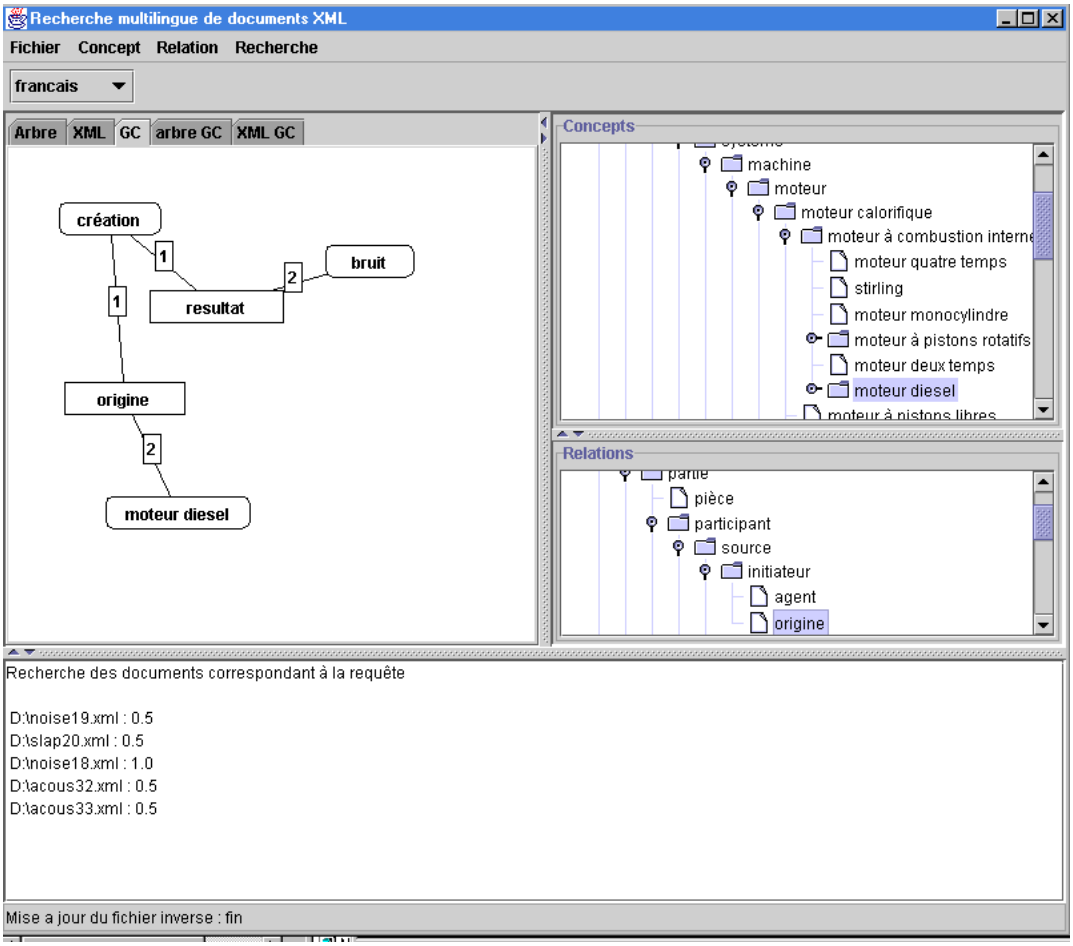


figure 1.19: Un exemple de recherche.

4.7. Conclusion :

La RI multilingue peut être vue comme un « mariage » entre la recherche d'information et la traduction automatique.
Tout d'abord, une présentation de la problématique générale de la recherche d'information multilingue, nous a permis de définir les particularités du CLIR. Ensuite, nous avons détaillé trois approches possibles au CLIR : la traduction de la requête, la traduction du corpus et la traduction du corpus et des requêtes dans un langage pivot.
Actuellement, l'approche la plus utilisée et donc la plus explorée est la traduction de la requête dans la langue du corpus.
Pour conclure, on remarque qu'il existe peu d'indexation adaptée au multilinguisme, c'est-à-dire que les entités d'indexation ne sont pas des termes trouvés dans le texte du document, mais appartiennent à un langage pivot non dépendant d'une seule langue.

Chapitre 2 : Les Ontologies

2.1. Introduction :

Un des enjeux actuels de la RI est de développer des systèmes capables d'intégrer plus de sémantique dans leurs traitements. L'objectif est double : « comprendre » les contenus des documents et « comprendre » le besoin de l'utilisateur pour pouvoir les mettre en relation.

Les ontologies sont utilisées pour représenter des descriptions partagées et plus ou moins formelles de domaine et ainsi ajouter une couche sémantique aux systèmes informatiques. C'est donc naturellement que des travaux sur l'intégration des ontologies dans les SRI se développent. Nous situons cette partie dans ce cadre là.

Nous commençons par les bases théoriques des ontologies. Nous évoquerons ensuite les différents composants constituant une ontologie, les différents types d'ontologies avant de donner les langages de représentation d'ontologies.

2.2. Bases théoriques :

2.2.1. Qu'est ce qu'une ontologie ?

La notion d'ontologie trouve son origine dans une branche de la philosophie traitant de la science de l'être. Cette discipline philosophique, initiée par Aristote, essaie de définir l'être ou, du moins, ce qui le caractérise. Le terme lui-même apparaît tardivement en 1692.
En philosophie, l'ontologie est l'étude de l'être en tant qu'être, c'est-à-dire l'étude des propriétés générales de ce qui existe [Amandine, 05].
En informatique, cette notion est apparue dans les années 90. Depuis, plusieurs définitions ont été proposées. D'après encyclopédie Wikipédia [PHAN, 05], une ontologie est un ensemble structuré de concepts. Les concepts sont organisés dans un graphe dont les relations peuvent être:

- Des relations sémantiques;
- Des relations de composition et d'héritage (au sens objet).

La structuration des concepts dans une ontologie permet de définir des termes les uns par rapport aux autres, chaque terme étant la représentation textuelle d'un concept. Par exemple, pour décrire les concepts entrant en jeu dans la conception de cartes électroniques, on pourrait définir l'ontologie (simplifiée ici) suivante:

- une *carte électronique* est un ensemble de *composants*
- Un *composant* peut être soit un *condensateur*, soit une *résistance*, soit une *puce*.
- Une *puce* peut être soit une *unité de mémoire*, soit une *unité de calcul* ;
- Une *carte électronique* qui contient une *unité de calcul* contient aussi au moins une *unité de mémoire*.

Simplement, on peut construire une ontologie à partir d'un corpus de textes. On va parcourir le texte à la recherche de termes récurrents ou définis par l'utilisateur, puis analysent la manière dont ces termes sont mis en relation dans le texte (par la grammaire, et par les concepts qu'ils recouvrent et dont une définition peut être trouvée dans un lexique fourni par l'utilisateur). Le résultat est une ontologie qui représente la connaissance globale que contient le corpus de texte dans le domaine d'application qu'il couvre.

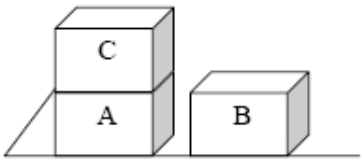
La définition la plus couramment citée est celle de T.Gruber : « an explicit specification of a conceptualization » [Gruber, 93]. Nous interpréterons cette définition comme «Une ontologie est une spécification formelle explicite d'une conceptualisation partagée »[Pierra, 02]. Dans cette définition, la conceptualisation signifie un modèle abstrait d'un certain aspect du monde. Les attributs "formelle" et "explicite" signifient qu'une ontologie permet une interprétation automatisée de la conceptualisation par la machine. Autrement dit, une ontologie définit un vocabulaire commun pour les chercheurs qui ont besoin de partager l'information dans un domaine. Les ontologies peuvent être utilisées par des personnes, des bases de données et des applications qui ont besoin de partager des informations sur un domaine.
Elles incluent des définitions, informations exploitables, des concepts élémentaires dans ce domaine et de leurs relations. Elles codent une connaissance dans un domaine et aussi une connaissance qui peut s'étendre sur plusieurs domaines. De cette façon, les ontologies permettent la réutilisation des connaissances.
Une ontologie regroupe ainsi les définitions d'un ensemble structuré de concepts. Les concepts sont organisés dans un graphe dont les relations peuvent être des relations sémantiques ou des relations de composition et d'héritage (au sens objet). La structuration des concepts dans une ontologie permet de définir des termes les uns par rapport aux autres, chaque terme étant la représentation textuelle d'un concept.
L'ontologie est donc la définition d'un domaine, une sorte de dictionnaire non linéaire et complexe. A la lecture d'une ontologie, la personne est censée comprendre tout le domaine concerné.
Ces définitions sont traitables par machine et partagées par une communauté de personnes. Elles doivent, en plus, être explicites, c'est-à-dire que toute la connaissance nécessaire à leur compréhension doit être spécifiée.

- Ontologie pour le Web :

Pour le Web, simplement, une ontologie est un document ou fichier qui définit de façon formelle les relations entre les termes. Ce type d'ontologie possédera une taxonomie et un ensemble de règles d'inférence. La taxonomie définit des classes d'objets et les relations entre eux.
Les classes, les sous classes et les relations entre les entités sont un très puissant outil pour utiliser sur le Web. On peut exprimer ces relations en attribuant des propriétés aux classes et en permettant à des sous classes d'hériter de leurs propriétés.
Par exemple: Une *adresse* peut être définie comme un type de *lieu*, et les *codes postaux* de ville peuvent être définis pour s'appliquer seulement à des *lieux* etc.. Les villes ont des sites Web et les *codes postaux* doivent être un type de *ville.* Donc, on peut associer un site Web au *code postal d'une ville*. (Ici, il n'y aucune base de données qui relie directement ce code à un site Web). Les règles d'inférence sont plus puissantes. Une ontologie peut exprimer la règle suivante: «Si un *code postal de ville* est associé à un *code d'état* et qu'une *adresse* utilise ce code de ville, alors cette *adresse* est associée au *code de l'état* « . Les ordinateurs peuvent manipuler les termes de façon plus efficace et significative.

2.2.2. Au-delà des définitions :

Pour mieux illustrer et exemplifier ces définitions nous pouvons réutiliser la situation bien connue des cubes sur une table.



Cube (x): L'entité X est un Parallélépipède à angles droits avec tous ses côtés de longueurs égales.

Table: un objet global, qui est un meuble composé d'un dessus plat horizontal déposé sur un ou plusieurs pieds.

Sur (Cube: X, Cube: Y / Table): Une relation dénotant qu'un cube X est sur un autre cube Y ou sur la table.

```
Cube (A)
Cube (B)
Cube (C)
Sur (A,Table)
Sur (C,A)
Sur (B,Table)
```

a- réalité b - Ontologie c – Décrire l'état de la chose

Figure 2.1 : exemple de description de situation

La figure 2.1 montre un schéma décrivant la vraie scène de trois cubes arrangés sur une table.
[Gandon, 02] propose un vocabulaire conceptuel (ontologie de jouet) pour parler de quelques aspects de cette réalité (certains d'entre eux sont ignorés, par exemple il n'y a aucun vocabulaire pour exprimer les dimensions des cubes). Enfin l'état de la question de la scène observée est décrit en utilisant les primitives de l'ontologie.

2.2.3. Les objectifs de l'ontologie :

On utilise l'ontologie dans différents domaines : la représentation d'informations et de connaissances, l'intégration des systèmes d'informations, la spécification des systèmes, etc. Mais aussi dans :
- **La communication :** L'ontologie ne permet jamais que deux mots différents possèdent la même sémantique.
- **L'interopérabilité :** L'ontologie peut être considérée comme un pont ou une passerelle entre les différents systèmes. "Elle sert à définir le format d'échange entre les systèmes."[INRIA, 01][Malik, 02].

Voici donc quelques raisons développer une ontologie:
- *Partager la compréhension commune de la structure de l'information entre les personnes ou les fabricants de logiciels*

Par exemple: On a certain nombre de sites Web contiennent de l'information médicale ou fournissent des services de ecommerce en médecine. Si ces sites partagent et publient tous la même ontologie, qui est à la base des termes qu'ils utilisent, alors les agents informatiques peuvent extraire et agréger l'information de ces différents sites. Les agents peuvent utiliser cette information agrégée pour pouvoir répondre aux interrogations des utilisateurs ou comme données d'entrées pour d'autres applications.
- *Permettre la réutilisation du savoir sur un domaine*
- *Expliciter ce qui est considéré comme implicite sur un domaine*
- *Analyser le savoir sur un domaine*

Le Web sémantique a besoin d'ontologies ayant un degré de structure significatif. C'est pour quoi on va définir des descriptions pour les concepts suivants:
• Les classes (de choses générales) dans les nombreux domaines d'intérêt;
• Les relations pouvant exister entre les choses;
• Les propriétés (ou les attributs) attachés à ces choses.

Autrement dit, une ontologie est un modèle d'organisation des connaissances dans un domaine donné. On trouvera dans l'ontologie les classes d'objets à organiser (personnes, étudiant, professeur, thèse...), les types d'attributs pouvant être attachés aux objets (référence, description, adresse, nom...) et les types de relations entre les objets (un objet "étudiant " peut être relié par une relation "supervisé par" à un objet de type "professeur"), etc...

2.2.4. Composants des ontologies :

D'après [Ahcene , 05] et selon Gomez-Pérez, les connaissances traduites par un e ontologie sont véhiculées à l'aide de cinq éléments [Gomez, 99] : Concepts ; Relations ; Fonctions ; Axiomes ; Instances.

- **Les concepts :** ils sont appelés aussi termes ou classes de l'ontologie, un concept est un constituant de la pensée (un principe, une idée, une notion abstraite) sémantiquement évaluable et communicable. Selon [Gomez, 99], ces concepts peuvent être classifiés selon plusieurs dimensions : 1)niveau d'abstraction (concret ou abstrait) ; 2) atomicité (élémentaire ou composée) ; 3) niveau de réalité (réel ou fictif). En résumé, un concept peut être tout ce qui peut être évoqué : description d'une tâche, d'une fonction, d'une action, d'une stratégie ou d'un processus de raisonnement, etc.
- **Les relations :** elles traduisent les associations existant entre les concepts présents dans le segment analysé de la réalité. Ces relations regroupent les associations suivantes : sous-classe-de (spécialisation, généralisation) ; partie-de (agrégation ou composition) ; associée-à ; instance-de ; est-un, etc. ces relations nous permettent d'apercevoir la structuration et l'interrelation des concepts, les uns par rapport aux autres. Les relations représentent un type d'interaction entre les notions d'un domaine. Elles sont formellement définies comme tout sous-ensemble d'un produit de n ensembles, c'est à dire R : C1*C2*……Cn.
- **Les fonctions :** sont des cas particuliers de relations dans lesquelles le nième élément de la relation est défini de manière unique à partir des n-1 premiers. Formellement, les fonctions sont définies ainsi : F :C1*C2*…..*Cn-1→Cn. Comme exemple de relation binaire, nous avons la fonction mère de.
- **Les axiomes :** pour structurer des phrases qui sont toujours vrais. Ils constituent des assertions, acceptées comme vraies, à propos des abstractions du domaine traduites par l'ontologie.
- **Les instances :** elles sont utilisées pour représenter des éléments.

2.2.5. Types d'ontologies :

Un critère pour la classification des ontologies est le contenu de la connaissance qu'elles représentent, c'est-à-dire le sujet de la conceptualisation [Guiraude, 02].

- **Les ontologies de domaine :** rassemblent les connaissances pour un domaine particulier (la médecine, la mécanique,...) et elles sont réutilisables dans un domaine donné. Elles fournissent le vocabulaire des concepts d'un domaine et les relations entre ces derniers, les activités de ce domaine (par exemple anesthésier, accoucher) ainsi que les théories et les principes de base de ce

domaine. Les ontologies de domaine ont pour avantage de permettre une normalisation des concepts dans le cadre du domaine considéré et dons, selon nous, de permettre une meilleure représentation de connaissance. Selon [Mizoguchi, 00][Ahcene , 05], l'ontologie de domaine caractérise la connaissance de domaine où la tâche est réalisée. Par exemple, dans le contexte du e_learning, le domaine est celui de formation.

- **Les ontologies applicatives :** contiennent toutes les définitions qui sont nécessaires pour modéliser la connaissance propre à l'élaboration d'une tâche particulière [Nathalie, 05]. Généralement, les ontologies d'application combinent des éléments d'ontologies de domaine et d'ontologies génériques choisies en fonction des méthodes spécifiques pour réaliser la tâche visée. Elles sont rarement réutilisables pour une autre application.
- **Les ontologies génériques :** ces ontologies expriment des conceptualisations valables dans différents domaines. Elles définissent des concepts considérés comme génériques à plusieurs domaines [Nathalie, 05].
- **Les ontologies de représentation de la connaissance :** regroupent les primitives de représentation utilisées afin de formaliser les connaissances. Elles permettent d'expliquer la conceptualisation sous-jacente aux formalismes de représentation [Davis, 93][Nathalie, 05]. Elles proposent un cadre de représentation sans émettre d'hypothèse sur le monde. On les désigne également comme ontologies abstraites ou de haut niveau parce qu'elles permettent de définir des concepts abstraits et peuvent être re-utilisées pour définir des concepts spécifiques. Un exemple d'ontologie de ce type est la Frame Ontology utilisée dans Ontolingua [Gruber, 93][Nathalie, 05], qui rassemble les primitives de représentation (classes, instances, cases, facettes, etc.) utilisées dans les langages à base de frame.

Nicola Guarino [Gua98][Guiraude, 02] distingue, lui les top-level ontologies à rapprocher des ontologies génériques de Van Heijst, les domaines and task ontologies décrivant respectivement de vocabulaire d'un domaine et les tâches et activités, et enfin les applications ontologies décrivant les concepts relatifs à la fois au domaine et aux tâches.

Deux types d'ontologies peuvent dés lors être distingués.

- Le premier regroupe les ontologies servant au stockage et à la mise en forme de connaissances d'un domaine particulier nécessaires à un système. Ce premier type, que nous nommons *ontologies de domaine* rassemble les ontologies de domaine et les ontologies applicatives de Van Heijst. Ces ontologies ont alors un rôle bien défini : porter les connaissances utiles et nécessaires et être intégrées à un système informatique donné. Ces ontologies peuvent être rapprochées des terminologies [FMJ01][Guiraude, 02], dés lors que toutes deux présentent un panorama d'un domaine de connaissance donnée. Les ontologies terminologies ou linguistiques spécifient les termes utilisés pour représenter la connaissance d'un domaine.
- Le deuxième type d'ontologies, rassemblant les ontologies génériques et les ontologies de représentation de Van Heijst et les top-level ontologies de Guarino n'a pas d'utilité en tant que tel comme élément d'un système informatique. Elles permettent de stocker et de formaliser des connaissances sur les connaissances. Ces ontologies peuvent servir à l'élaboration d'ontologies de domaine. Ainsi une ontologie générique va expliquer les concepts d'un domaine particulier, décrirez les éléments qui le composent.

2.2.6. Les déférents modes de représentation des ontologies :

L'objectif de nos travaux étant la prise en compte par le SRI de ressources conceptuelles. Les langages dédiés aux ontologies sont principalement issus des formalismes liés aux réseaux sémantiques, les graphes conceptuels et les frames.

2.2.6.1- Réseaux sémantiques :

Un réseau sémantique est une représentation graphique d'une conceptualisation d'une (ou plusieurs) connaissance humaine [Quillian 1968][Nathalie, 05]. Il est représenté sous la forme d'un graphe orienté et étiqueté ou plus précisément un multigraphe car deux nœuds du graphe peuvent être reliés par plusieurs arcs. Il est constitué d'un ensemble de nœuds typés, dénotant des concepts du domaine modélisé, et d'arcs orientés étiquetés représentant les relations sémantiques entre les concepts. Ainsi un concept est décrit par les autres concepts du réseau en lien avec lui, comme l'illustre la figure 2.2 certains nœuds correspondent intuitivement mieux à des classes d'objets qu'à des individus. Représenter l'appartenance à une classe nécessite une relation d'appartenance ; c'est pourquoi les réseaux possèdent un nom réservé d'étiquette pour cette relation, parfois nommé ''sorte de''. Mais cette relation d'appartenance à une classe suppose que l'on sait différencier les nœuds qui représentent des classes de ceux qui dénotent des individus. De nombreux formalismes de représentation des connaissances dérivés des réseaux sémantiques imposent de noter différemment les deux types de nœuds (par exemple KL-ONE [BRAC85][Roussey, 01]. La relation d'appartenance à une classe permet de rapporter les connaissances de la classe sur un individu. Cette opération est appelée inférence par héritage.

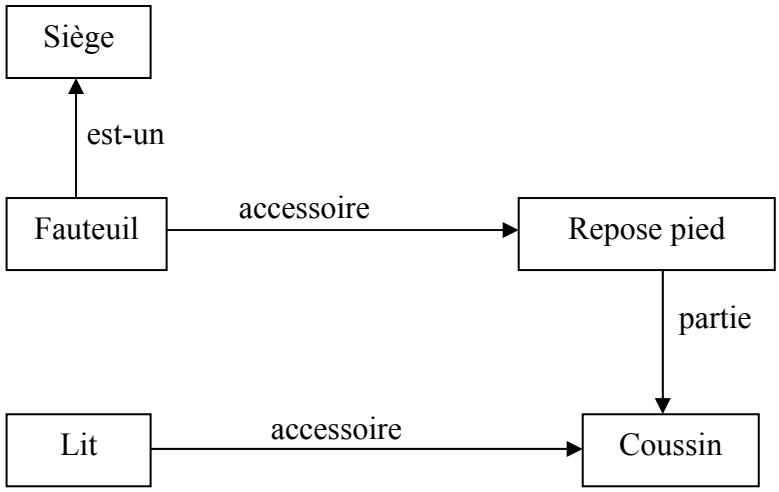


Figure 2.2 : Un exemple de réseau sémantique décrivant du mobilier.

Le filtrage est un mécanisme de recherche de tous les sous-graphes du graphe qui ont une structure commune avec un graphe cible. Dans le cadre d'un système de recherche d'information, la requête est modélisée sous forme d'un réseau sémantique dans lequel les connaissances inconnues sont exprimées par des variables. Le système met en correspondances ce réseau requête et une partie du réseau contenant l'ensemble des connaissances afin de déduire les valeurs de ces variables. Il peut faire appel à l'héritage pour récupérer des informations dés concepts généraux et les utiliser dans les nœuds plus spécifiques.

Cependant, ce type de définition ne concerne que la structure du graphe et ne permet pas d'ajouter de l'information sémantique. De nombreuses études [Woods 1975], [Brachman 1977][Nathalie, 05] ont montré que ce type de graphe manque de précision sémantique et mène à des confusions entre les relations et aussi entre les classes et individus. Elles ont mené à la définition de nouveaux formalismes tels que les frames , les logiques de description et les graphes conceptuels.

2.2.6.2. Les graphes conceptuels :

Ont été proposés par Sowa en 1984 [Sowa, 84] et utilisent une notion à base de graphe. Pour présenter les graphes conceptuels, Roussey [Roussey, 01] a été inspiré de l'exposé très rigoureux de [CHEI, 92]. Tout d'abord, il définit le support qui va régir l'ensemble des graphes conceptuels portant sur un même domaine de la connaissance. Un support se compose :

- d'une hiérarchie Tc de type de concepts organisés par une relation de spécialisation notée $\leq$, dans un treillis. Un exemple de treillis de type de concepts peut contenir le type 'Automobile' qui spécialise le type 'Véhicule' et inversement 'Véhicule' généralise 'Automobile' ('Automobile' $\leq$'Véhicule'). Un treillis possède un plus grand élément appelé le type universel, noté T et un plus petit élément, le type absurde, noté $\perp$.
- D'un ensemble Tr de type relation composé de plusieurs hiérarchies de types de relations de même arité (ayant le même nombre d'arguments). Chaque hiérarchie est organisée en treillis par la relation de spécialisation $\leq$. Les arguments de chaque type de relations r_i de Tr sont numérotés, comme l'indique la figure 2.3, et ils obéissent à des contraintes de typages représentées dans la signature de r_i .
- D'un ensemble M de marqueurs. Un marqueur identifie un individu de la base de connaissances. On dispose également d'un marqueur générique noté * qui représente un individu non spécifié.
- D'une relation de conformité entre les marqueurs et la hiérarchie des concepts. Cette relation doit obéir à certaines contraintes : tout marqueur est conforme au type universel et aucun au type absurde. Si un marqueur est conforme à un type t, il est aussi conforme à tous les types généralisant t ; si un marqueur est conforme à deux types t et t', il est aussi conforme au type spécialisant t et t' ($t \cap t'$).

Un graphe conceptuel est un multigraphe, composé de deux sortes de nœuds : les nœuds concepts, aussi appelés sommets concepts ou plus sommairement concepts, et les nœuds relations ou relations. Chacun de ces nœuds a une étiquette. Un nœud concept est étiqueté par un type correspondant à une classe sémantique, et un marqueur précisant une instance particulière de classe. Par exemple, le nœud concept étiqueté par 'Automobile : *' représente une automobile en générale. Ce genre de nœud est qualifié de concept générique. Le nœud concept étiqueté par 'Automobile : 6793 SY 69' représente une voiture particulière dont le numéro d'immatriculation est 6793 SY 69. ce nœud est qualifié de concept individuel. Les relations spécifient les rapports entre les concepts. Dans les graphes conceptuels, un concept est appelé un **argument** de la relation. Les nœuds relations sont aussi étiquetés par un type. Par exemple, la figure 2.3 présente le graphe conceptuel signifiant qu'une automobile est composée d'un moteur.

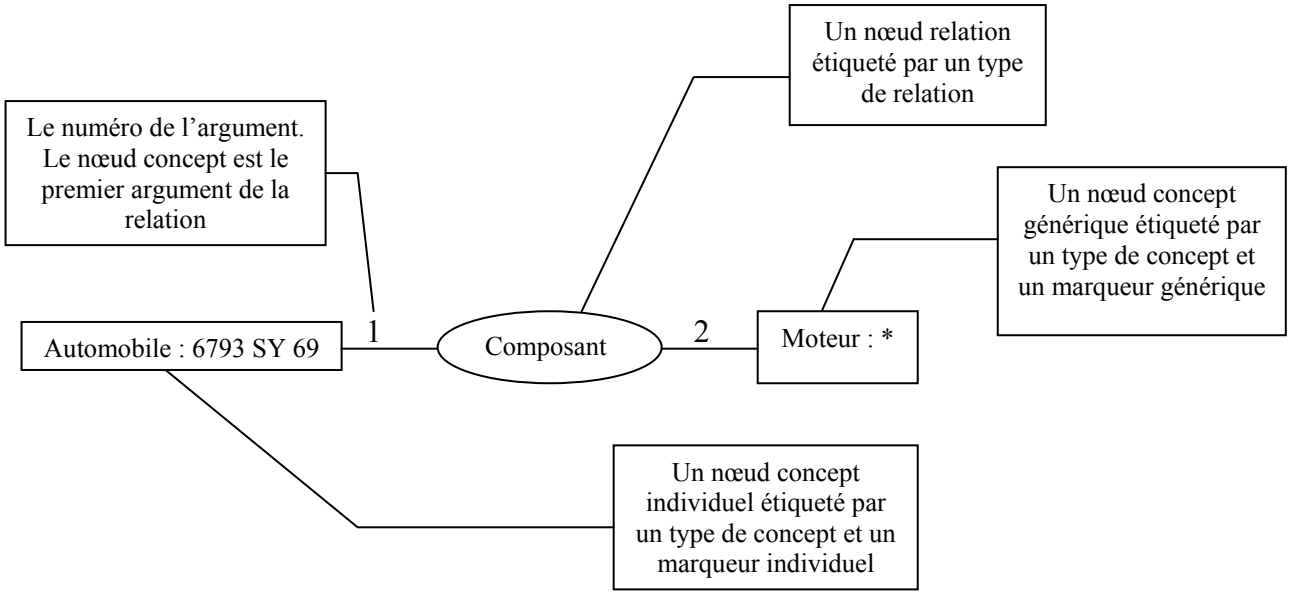


Figure 2.3: Le graphe conceptuel : '' une automobile est composée d'un moteur''.

Pour la recherche d'information, la relation de spécialisation présente un intérêt majeur car elle contribue à la comparaison de graphes. En effet, pour les comparer, Sowa a proposé un opérateur de projection qui s'appuie sur l'existence d'une relation de spécialisation entre deux graphes. Comme le montre la figure 2.4, il existe une projection d'un graphe H dans un graphe G si le graphe G reprend de manière plus spécifique l'ensemble des concepts présents dans le graphe H. dans ce cas, G est dit une spécification de H, $G \leq H$.

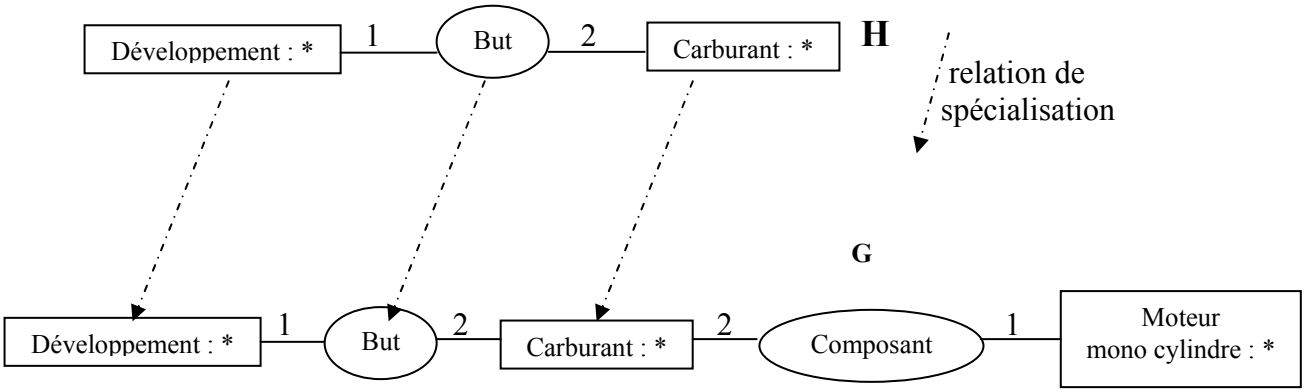


Figure 2.4 : Un exemple de projection

De plus, Sowa a défini un opérateur ϕ qui permet de transformer un graphe conceptuel en une formule logique du premier ordre et il a montré que l'existence d'une projection d'un graphe H dans un graphe G était conditionnée par l'implication de leurs formules logiques associées : $G \leq H \Rightarrow \phi(G) \rightarrow \phi(H)$. Cette implication fait des graphes de Sowa un modèle adapté pour construire des systèmes de recherche d'information tels que l'a préconisé Van Rijsbergen. En effet, s'il existe une projection d'un graphe H dans un graphe G représentant respectivement une requête et l'index d'un document, elle se traduira par l'implication $\phi(G) \rightarrow \phi(H)$ qui rend le document pertinent pour la requête.

Donc les graphes conceptuels peuvent être vus comme des schémas permettant de représenter graphiquement des formules logiques, ou bien des schémas sans contraintes, servant juste d'interface « graphique » à la représentation de formules ou bien comme des graphes munis d'opérations de graphes permettant le raisonnement et leur manipulation en s'appuyant sur la théorie des graphes. Les graphes conceptuels ont été utilisés dans les systèmes d'information pour la représentation de requêtes et de documents dans [Guarino 1999][Nathalie, 05]. Ils sont élaborés manuellement et une ressource lexicale (WordNet) est utilisée pour les mettre en correspondance avec les requêtes de l'utilisateur.

2.2.6.3. Les frames :

Dans les langages de frames, les connaissances sont regroupées en paquets. Ainsi les relations d'un concept du domaine avec les autres concepts font partie de la description de ce concept. Dans les réseaux sémantiques, il existe une unité sémantique dénotant un concept (le représentant), et un graphe composé d'autres unités sémantiques décrit la définition de ce concept. Dans un frame, l'unité sémantique, dénotant le concept, est confondue avec sa description, c'est à dire qu'une seule unité sémantique contient toute la description du concept et est aussi utilisée pour représenter ce concept.
En 1975, Minsky propose le premier formalisme informatique à partir des idées précédentes [MINSKY, 75][Roussey, 01]. Le frame correspond à une structure dynamique représentant des situations prototypages. Schank, qui s'intéresse à la représentation de séquence d'évènements et la compréhension d'histoires, aboutit quasiment au même résultat avec la théorie des scripts [SCHANK, 77][Roussey, 01].
Un frame est un prototype, c'est à dire un objet typique d'une famille, représentant idéalement cette famille. Le frame contient donc des informations générales valides pour tous les membres de la famille, ainsi que des informations spécifiques à certains membres.
Les frames sont représentés comme étant une structure de données capable de représenter des objets structurés.
Un frame est composé d'un ensemble d'attributs ou slots correspondant aux différentes propriétés du prototype. Un attribut est décrit par un certain nombre de facettes ou rôles possédant des valeurs. Par exemple, une facette définit le type de l'attribut, une autre la valeur courante ou par défaut de cet attribut. Les facettes procédurales permettent d'associer aux attributs des procédures appelées réflexes (par exemple, une procédure peut être déclenchée pour calculer la valeur d'un attribut).
Une des notions centrales des langages de frame est la notion de spécialisation, car elle permet non seulement l'ordonnancement de ces frames dans une hiérarchie mais l'héritage des attributs et de leurs valeurs. Ainsi la description d'un frame peut être incomplète, les couples attribut-valeur hérités ne sont pas recopiés dans le frame. Un frame peut être spécialisé par enrichissement, en le dotant de nouveaux couples attributs valeurs ou par substitution, en masquant certaines facettes des couples attribut-valeur hérités.
Dans le langage de frames proposé par Minsky, la notion d'instanciation n'existe pas, la hiérarchie des frames permet d'affiner successivement les descriptions jusqu'à la spécialisation d'un concept ne pouvant plus être spécialisées. Une évolution des frames de Minsky a permis de distinguer deux types de frames : les frames classes et les frames instances. Les frames classes représentent les catégories d'objets du monde modélisé alors que les frames instances représentent des individus particuliers. La

hiérarchie structurant les frames classes les uns par rapport aux autres, par la relation de spécialisation, est appelée une **taxinomie**. En plus de cette relation de spécialisation, des systèmes de frames proposent implicitement ou explicitement la relation d'instanciation (intitulée est-un dans les cas où elle existe de façon explicite), qui relie une instance à sa classe d'appartenance, comme le montre la figure suivante :

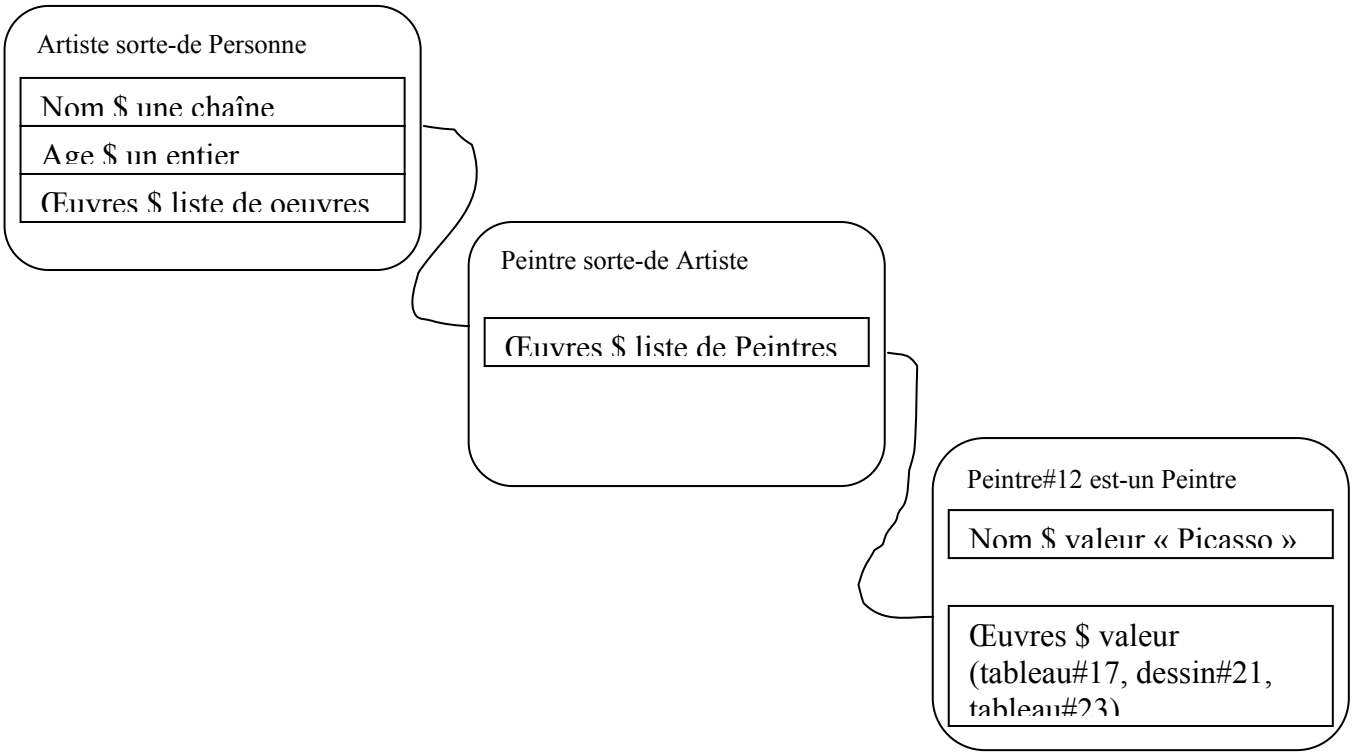


Figure 2.5 : Exemple de spécialisation de frames. Artiste et Peintre sont des classes, alors que Peintre#12 est une instance.

Un système de frame sert à comparer des objets que l'on veut reconnaître ou classer. Donc deux mécanismes de raisonnement sont disponibles :

- Le filtrage consiste à rechercher parmi un ensemble de frames ceux qui correspondent à des critères donnés. Il repose sur des mécanismes d'héritage et d'appariement.
- La classification consiste à intégrer un nouveau frame dans une hiérarchie établie. Le réajustement correspond à la modification de la position d'un frame mal placé.

Dans le cadre de la recherche d'information, un document est représenté par un frame, qui contient aussi une description de son contenu. La requête, elle, spécifie certaines caractéristiques que doit remplir le frame document. La fonction de comparaison effectue donc un filtrage sur l'ensemble des frames pour identifier tous les frames documents correspondant aux critères de la requête.

2.2.6.4. Les logiques de description :

Les logiques de description, que l'on appelle également logiques terminologiques, représentent le formalisme de représentation des connaissances le plus répandu actuellement. Ces logiques sont issues de la théorie des frames à laquelle s'ajoutent des principes des réseaux sémantiques. KL-ONE [Brachman, 85][Roussey, 01],

développé en 1978, est considéré comme le précurseur des logiques de description. Ce langage a fortement influencé le domaine de la représentation des connaissances, si bien que l'on parle maintenant de la famille ''KL-ONE'' regroupant tous ses descendants. Par exemple, nous pouvons citer les langages KRYPTON, LOOM, etc.
Dans ces travaux sur la représentation des connaissances, R.Brachman propose la séparation des connaissances en plusieurs niveaux [Brachman, 77]. De là sont nées les notions de T-Box et A-Box, que l'on retrouve dans la plupart des logiques de description :

- La T-Box (Terminological Box) permet de décrire les concepts en fonction d'autres concepts à partir de relations et de contraintes sur ces relations. Elle renferme les connaissances terminologiques.
- La A-Box (Assertional Box) correspond au niveau factuel ou niveau des assertions, et est réservée à la description et la manipulation des individus.

Les logiques de description comprennent trois composantes formelles : le **concept**, le **rôle** et **l'individu**. Le concept représente un ensemble ou une classe d'individus. Un concept correspond à la conjonction de rôles qui expriment les relations existantes entre celui-ci et les autres concepts. Des restrictions ou des cardinalités peuvent être affectées à chaque rôle, un peu à la manière des facettes associées à un attribut dans un frame. L'individu correspond à une entité particulière, une instance du concept. Concepts et rôles relèvent de la T-Box, alors que les individus sont en général défins dans la A-Box.
Au niveau terminologique, on distingue deux types de concepts : les concepts primitifs et les concepts défins. Un concept primitif possède une description incomplète, correspondant à des conditions nécessaires. Les concepts primitifs servent à construire les concepts définis. Un concept défini possède une description complètement spécifiée, qui correspond à des conditions nécessaires et suffisantes exprimées par des rôles. Dans l'exemple suivant, les concepts sont écrits en majuscules et les rôles en minuscule. Les concepts primitifs sont introduits par le symbole :< et les concepts définis sont introduits par le symbole :=.

PERSONNE :< TOP
ENSEMBLE :< TOP
HOMME :< PERSONNE
FEMME :< (and PERSONNE (not HOMME))
ministre :< top-role
premier-ministre :< ministre
GOUVERNEMENT := (and ENSEMBLE (some ministre)(all ministre PERSONNE))
GOUVERNEMENT-MIXTE := (and GOUVERNEMENT
(some (range ministre FEMME))

Les concepts du domaine sont organisés en une hiérarchie par la relation de subsomption. La relation de subsomption est définie de plusieurs manières :

- Définition extensionnelle : un concept A subsume un concept B si l'ensemble des individus dénotés par A contient l'ensemble des individus dénotés par B [Roussey, 01]. Par exemple, le concept mammifère subsume le concept dauphin.
- Définition intensionnelle : un concept A subsume un concept B si tout individu décrit par B l'est aussi par A ; autrement dit si l'ensemble des propriétés spécifiées par A. par exemple des propriétés associées au concept dauphin comprend l'ensemble des propriétés associées au concept mammifère.

- Définition logique : un concept A subsume un concept B, si être un individu décrit par B implique être un individu décrit par A.

La classification de concepts constitue le mécanisme d'inférence le plus important des logiques de description. Il s'agit d'un processus permettant de déterminer la position d'un concept donné dans la hiérarchie de subsomption. Plus précisément la classification consiste à placer un concept dans la hiérarchie en spécifiant les concepts qui subsument et les concepts qui sont subsumés par lui. La classification est mise en œuvre grâce à un programme spécialisé appelé le classifieur [SCHMOLZE, 83][Roussey, 01].
Signalons que le processus de classification est utilisé dans les systèmes de recherche d'information. L'idée de base consiste à reprendre le modèle logique de RI en représentant les requêtes par des concepts et les documents par des individus plutôt que par des expressions logiques. L'implication $D \rightarrow Q$ s'énonce par conséquent comme suit : l'individu D est une instance du concept Q, ou la requête Q subsume le concept décrivant le document D. Ainsi les documents jugés pertinents par le système correspondent aux instances du concept représentant la requête et de ses subsumés [OUNIS, 95][Roussey, 01].

Les logiques de description sont plus flexible que les frames et reposent sur une sémantique et une syntaxe rigoureuse. Elle est utilisable en RI car elle permet de traiter des données erronées ou incomplètes tout en offrant la possibilité d'ordonner hiérarchiquement les données [Nathalie, 05]. Cependant, elles nécessitent l'élaboration manuelle de ressources de connaissances formalisées à partir de cette logique.

2.3. Langages de spécification d'ontologie pour le Web sémantique :

Le langage de spécification est l'élément central sur lequel repose l'ontologie.
Bien qu'aujourd'hui, le Web sémantique soit un domaine établi, avec ses langages associés à peu près définis, il s'inspire d'un certain nombre de projets ou initiatives.
C'est pourquoi, nous allons maintenant en présenter certains :

2.3.1. SHOE : (Simple HTML Ontology Extension) :

Etait une extension de HTML ayant pour but d'incorporer des aspects sémantiques dans les pages Web. C'est une petite prolongation au HTML qui permet à des auteurs de page Web d'annoter leurs documents de Web avec la connaissance compréhensible par une machine et de rend le vrai logiciel intelligent d'agent sur le Web possible [SHOE, 06].
Le consortium W3C a proposé Extensible Markup Language (XML) [BRAY et al, 98] [PHAN, 05] qui nous permet de créer les ensembles des étiquettes pour adapter aux besoins du client dans ses documents. Puis pour montrer cette information dans quelque format approprié, on peut utiliser StyleSheet. SHOE est basé sur XML: XML permet à SHOE d'être ajoutée aux pages Web et SHOE ajoute à XML une manière standard pour exprimer la sémantique dans un contexte indiqué. RDF est un autre travail de W3C. RDF indique les réseaux sémantiques d'information sur des pages Web, mais il n'a pas la possibilité déductive et il est limité aux relations binaires entre des objets. On va le parler dans la partie suivante. Il y a beaucoup d'autres projets qui emploient des ontologies pour le Web sémantique.

D'abord, on a une question : «*est-ce qu'on peut avoir un robot intelligent d'agent qui va utiliser un système de traitement de langage pour lire les phrases et avoir la capacité de dire les significations de celles ci dans une page Web?* « La réponse est non ! Pourquoi ? Parce qu'on sait que le traitement de langage naturel a un chemin très long d'aller. On a plein des langages dans le monde et plein des obstacles pour surmonter. En plus, les Web pages normales ont été non seulement écrits dans une langue lisible pour l'homme mais dans une disposition humain-vision-orientée parce que HTML est pour montrer des données pour que les humains lisent. Nous sommes les seules qui peuvent lire les connaissances dans une page Web grâce à la présentation sur formes des tables, des graphiques et des «frame » . C'est à dire nous comprenons visuellement. Mais les agents intelligents ne sont pas humains. Donc, c'est très difficile pour lui de comprendre le Web page par la lecture. Avec SHOE, on peut élimine ce problème en ajoutant la connaissance que les agents intelligents peuvent réellement lire dans les pages Web.
SHOE est une langue HTML basée de représentation de la connaissance. Il va ajouter les étiquettes nécessaires pour inclure des données sémantiques arbitraires dans des pages Web. Des étiquettes de SHOE sont divisées en deux catégories. Premièrement, il y a des étiquettes pour construire des ontologies. Ici, les ontologies sont des ensembles de règles qui définissent quels genres de documents que SHOE peut d'affirmer et ce que signifient ces affirmations. Par exemple, une ontologie dans SHOE pourrait indiquer qu'un document qui déclare que l'entité de quelques données est un "étudiant à Boumerdes ", et que s'il ainsi, que ce "étudiant" est autorisé pour avoir un "ordinateur". Deuxièmement, il y a des étiquettes pour annoter des documents sur Web en utilisant un ou plusieurs ontologies, pour déclarer des entités de données, et pour faire des affirmations au sujet de ces entités selon les règles déjà établi par les ontologies. Par exemple, un document de SHOE utilise l'ontologie de SHOE ci-dessus pourrait alors déclarer qu'il est tout concernant d'un étudiant qui est en train d'utiliser la machine "0046E" par exemple.

SHOE contient deux types de balises, les balises définissant les ontologies et celles qui déclarent les instances. Une ontologie définit les éléments valides pour décrire les instances. Ces éléments sont :

- Les catégories organisées en hiérarchie par la relation de spécialisation.
- Les relations qui peuvent exister entre instances ou qui décrivent les propriétés d'une instance.
- Les règles permettant de renommer des éléments empruntés à d'autres ontologies.
- Les règles d'inférences définissant des inférences sous forme de clauses de Horn.
- Les constantes.
- Les types de données.

Pour SHOE, une page web est considérée comme une instance. Afin de spécifier les relations entre instances contenues dans une page, chaque page web est associée à une ontologie préalablement définie. Les pages web définissent en fait des croyances et non des connaissances absolues ; par conséquent, un utilisateur peut spécifier son degré de confiance pour une source information.
Plusieurs technologies ont été mises en œuvre pour démonter l'efficacité de ce langage :

- Un outil d'annotation permet aux auteurs de documents HTML d'enrichir leurs pages en incluant des connaissances exprimées à l'aide du langage SHOE.
- Un robot intitulé **Exposé** cherche sur le web des pages annotées en langage SHOE.

Ensuite, exposé analyse les pages web pour identifier les connaissances contenues dans les pages web. Ces connaissances, qui peuvent être des ontologies ou des instances, sont enregistrées dans un système à base de frames, intitulé PARKA [EVETT, 93][Roussey, 01].

- Un outil interrogant la base de connaissances construite par Exposé et visualisant les pages web résultats. Les requêtes spécifient les attributs ou propriétés que doit avoir un frame instance. La page web identifiant le frame résultat est ensuite affichée.

SHOE fournit un mécanisme permettant de modifier des ontologies déjà existantes en spécialisant par exemple sa hiérarchie, ce qui veut dire que des parties des ontologies sont dispersées sur le web. Cette dispersion peut entraîner une redondance d'information et aussi une absence d'information. En effet, un utilisateur ne dispose peut être pas de toutes les connaissances disponibles sur un domaine avant d'annoter sa page ou d'interroger la base de connaissances.

2.3.2. Ontobroker : (On-2-broker),

C'est un projet qui emploie des ontologies pour le Web sémantique. Ce projet, comme SHOE, est inclus dans le HTML. Bien que la syntaxe de cette langue soit plus compacte, il ne peut pas être compris de façon assez facile comme SHOE. En outre, Ontobroker n'a pas un mécanisme pour que les pages puissent utiliser plusieurs d'ontologies. On n'a que une manière de le découvrir si l'on est un des membres de la communauté [PHAN, 05].

Il utilise les F-Logic (Frame Logic), comme formalisme de représentation pour définir les ontologies et pour exprimer les annotations de documents [On2broker, 04].

Donc, ce système a pour but de représenter formellement les connaissances contenues dans les documents HTML pour pouvoir interroger de manière précise ces connaissances stockées dans une base de connaissances. Dans cette approche, le web est considéré comme une base de connaissances distribuée, non structurée, et même non exprimée car implicite. Leur but est de réutiliser et partager ces connaissances. Tout d'abord, un groupe d'utilisateur, appelé **ontogroup**, voulant mettre en commun leurs connaissances et leurs informations s'accorde sur la définition d'une ontologie.

L'ontologie définit non seulement leur vocabulaire commun mais stocke et organise les connaissances à l'aide du langage Frame Logic [KIFER, 95][Roussey, 01] sous forme de frames (les classes=[classe, attribut, type de valeurs] et les instances = [instance, attribut, valeur]) et de règles. L'ontologie est la représentation formelle consensuelle d'un point de vue sur le domaine partagé par les acteurs dans ce domaine.

Dans ce système, les documents sont considérés comme des instances de classe. Les connaissances sont explicitement représentées dans les pages HTML. Ainsi, à une partie de texte, est ajoutée la description formelle de sa sémantique. Les connaissances exprimées sont les valeurs des attributs de l'instance considérée (la page web). Le texte des documents ou les liens peuvent être réutilisés comme valeur d'un attribut. L'explication de ces connaissances peut être réutilisée comme valeur d'un attribut. L'explication de ces connaissances peut être faite automatiquement dans

le cas de document HTML fortement structuré à l'aide d'un wrapper, un médiateur entre la structure du document et l'ontologie ; ou manuellement en annotant le document. Un langage d'annotation a été défini. Il s'agit d'une extension des ancres HTML définissant un nouvel attribut, intitulé onto. Le fait d'expliquer les connaissances prés de leur source d'information dans le document lui-même, permet de faciliter la maintenance de la base de connaissances. En effet, plus besoin de gérer en parallèle deux sources d'information (le document et les connaissances associées), seule la mise à jour du document est nécessaire.

Ontobroker se décompose en trois éléments :

1- Un outil d'interrogation de la base de connaissances. Les requêtes sont des expressions du langage Frame Logic. Les requêtes sont une conjonction d'expressions du type : une instance O d'une classe C a pour attribut A, la valeur V. les variables O, C, A, V peuvent être remplacées par des constantes ou des expressions. Une interface sous forme de frame a été développée pour générer des requêtes de manière plus conviviale. Le résultat de ces requêtes sont des valeurs qui ne sont pas forcement des références à des documents. Par exemple, on peut rechercher tous les mots clés d'une équipe de recherche.
2- Un moteur d'inférence. Ce moteur traduit les expressions Frame Logique en formule logique du premier ordre pour pouvoir faire des inférences. Les inférences permettent d'améliorer la cohérence de l'ontologie. Par exemple, si un chercheur a un article dans sa liste de publication, alors cet article doit avoir ce chercheur dans la liste de ses auteurs.
3- Un webcrawler parcourant le web pour intégrer de nouvelles connaissances. Celle-ci étant stockée dans la même base.

Pour rendre le système Ontobroker compatible avec de nouvelles approches, un traducteur RDF a été développé pour représenter les connaissances liées au document sous forme d'une description RDF.

Ontobroker est à la base de deux projets : **Knowledge Acquisition initiative** $(KA)^2$ [BENJ98b][Roussey, 01] et **Ontoknowledge** [On2Knowledge, 02]. Le projet $(KA)^2$ développe une ontologie propre aux chercheurs de la communauté de l'acquisition des connaissances. Cette ontologie est utilisée pour stocker toutes les connaissances relatives aux laboratoires, projets, chercheurs, publications du domaine. Elle est utilisée en exemple dans Ontobroker. Ontoknowledge est un projet soutenu par la commission européenne visant à regrouper des chercheurs travaillant sur le web sémantique. Un des objectifs de ce projet consiste à définir un nouveau langage de représentation des ontologies **Ontology Interchange Layer** (OIL) [FENSEL, 00] [Roussey, 01] basé sur les schémas RDF.

En conclusion, Ontobroker n'est pas un outil de recherche documentaire, mais un outil de recherche de connaissances. Son but n'est pas de représenter le contenu des documents pour améliorer la recherche de document, mais de représenter les connaissances décrites dans les documents pour améliorer la recherche de connaissances.

2.3.3. Ontoseek :

C'est un système de coopération pour des agents intelligents [OntoSeek, 04].

Il est développé par Guarino et son équipe [GUARINO, 99][Roussey, 01], est un système de recherche de pages web utilisant le modèle des GC de Sowa. C'est-à-dire que le contenu des pages web et les requêtes sont représentés sous forme de graphes conceptuels. La fonction de comparaison est basée sur l'opérateur de projection de

Sowa et recherche donc les spécialisations des nœuds de la requête. Ontoseek exploite la base lexicale Wordnet [Roussey, 01] pour lever les ambiguïtés des termes utilisés comme étiquette des nœuds des graphes. Ainsi, étiquettes des nœuds ne sont plus des termes mais des synsets. Cette extension du modèle des GC de Sowa a été intitulée **Graphes Conceptuels Lexicaux** (GCL). Une des applications de ce système a été de travailler sur l'annuaire des pages jaunes. Les graphes indexant ces pages web ainsi que les requêtes sont construits manuellement. L'utilisateur peut choisir les composants de son graphe en naviguant dans l'ontologie ou, à partir d'un terme, Ontoseek lui présente différents synsets et l'utilisateur sélectionne celui qui correspondant à son besoin. Des heuristiques sont utilisées dans Wordnet sont insuffisantes pour déterminer les graphes non valides, par exemple le graphe [manger]-(patient)→[maison] est jugé valide par Ontoseek. En effet, Wordnet est une base lexicale et non une base de connaissances. Ce projet a montré l'intérêt d'utiliser une base lexicale pour lever les ambiguïtés terminologiques et ainsi améliorer la précision et le rappel du système de recherche.

2.3.4. Webkb :

C'est un projet mondial de base de connaissance qui emploie des ontologies pour le Web sémantique et l'étude de machine pour essayer de classifier automatique des pages Web [PHAN, 05] .
C'est un ensemble d'outils basé sur le Web permettant à ses utilisateurs de représenter la connaissance pour annoter des ressources et proposer des mécanismes de récupération en utilisant des graphiques conceptuels [WebKB, 06].
Il permet à ses utilisateurs de stoker, d'organiser et de retrouver des connaissances formalisées à l'aide du modèle des graphes conceptuels [Roussey, 01]. Ces outils ne sont pas dédiés à la recherche documentaire car leur but n'est pas de retrouver des documents mais de retrouver à partir d'une requête l'ensemble des connaissances répondant à cette requête, stockées dans une base de connaissances. Plus précisément, WebKB propose une série d'outils pour l'acquisition de connaissances, dont le but est de créer une base de connaissances illustrée par une documentation (l'ensemble des textes au format HTML) qui ont permis de valider les connaissances de la base. Les connaissances et les parties de documents associées sont touts les deux représentées sous forme d'Elément de Document (ED). Un ED doit être accessible par le web, en utilisant par exemple son URL. De plus, un ED est caractérisé par un contexte, constitué par son créateur et sa date de création. Les ED sont liés les uns aux autres par différents liens :

- Un lien hypertexte permet d'indexer une ED documentation à une ED connaissance représentant les connaissances déduites de l'ED documentation.
- Des liens sémantiques entre ED contenant des connaissances formelles, par exemple des liens de spécification entre deux graphes conceptuels.

WebKB comprend :

- Un outil de recherche sur la base de connaissances. Ainsi à partir d'une requête sous forme de graphe conceptuel sont retrouvés tous les graphes de la base répondant en partie à cette requête. L'opération de recherche a été élargie, et ne correspond plus uniquement à une projection au sens de Sowa. En conséquence, le graphe réponse peut être spécialisation ou une généralisation d'un sous-graphe du graphe requête. De plus, d'autres relations sémantiques que la relation de spécialisation entre types sont prises en compte

pour établir la pertinence d'un graphe par rapport à la requête, par exemple la relation '' partie de''. De plus, plusieurs contraintes sur la recherche peuvent être spécifiées, entre autres l'affichage des ED documentations associées aux ED connaissances retrouvés. La base de connaissances est chargée directement à partir des documents HTML. En effet, WebKB permet d'annoter les documents HTML à l'aide de commandes permettant de créer, modifier les connaissances de la bases [Roussey,01].

- Un outil d'indexation des ED permettant d'indexer des ED documentation par des ED connaissance et d'associer des ED connaissances par des liens sémantiques. Par exemple, les liens expriment la relation de subsomption ''sort de'' ou la relation d'appartenance ''partie de''.
- Un éditeur de connaissances permettant de créer des graphes conceptuels. Ces graphes peuvent être représentés à l'aide de plusieurs langages : le langage défini par Sowa, un langage proche du langage naturel intitulé Formalized English, le langage Frame-CG qui permet une représentation des graphes sous forme de frames. Il est intéressant de voir que WebKB a pour but de représenter le plus précisément et avec le moins d'ambiguïté possible les connaissances, c'est pourquoi le langage vde Sowa a été amélioré par l'ajout de plusieurs quantificateurs (tout, un, quelque, 96%, ect.).
- Un navigateur de hiérarchie permettant de rechercher différentes catégories. Une catégorie pouvant être un type de concepts, un type de relation, un marqueur. Dans WebKB les termes ne sont pas les identifiants des catégories, un identifiant de catégorie est unique, il se compose de la concaténation du nom du créateur de la catégorie, suivi par le terme le plus représentatif de la catégorie et si besoin est d'un numéro. Ainsi une différence est faite entre le niveau lexical (terminologie) et le niveau conceptuel. WebKB, afin de guider le cogniticien dans sa phase de modélisation du domaine, propose plusieurs hiérarchies de base (hiérarchie de concept, de relation) construites à partir de la base lexicale Wordnet [Roussey, 01].

Par le fait que ces outils intègrent à l'intérieur des documents HTML des commandes permettant de construire des connaissances, ce projet est intéressant pour le web sémantique. Par contre, on ne peut pas dire qu'ils soient destinés à la recherche documentaire sur le web, car les requêtes portent sur des connaissances précises et non sur le contenu global des documents. WebKB est un projet facilitant la construction de bases de connaissances illustrée par des parties de document. Ces bases de connaissances pourront être ensuite réutilisées pour rechercher des documents. Ces bases de connaissances pourront être ensuite réutilisées pour rechercher des documents sur le web.

2.3.5. CONCERTO :

(CONCEptual indexing, querying and ReTrieval Of digital documents), il est consacré à l'annotation et à la récupération de documents textuels dans les domaines de la biologie et de la publication [Concerto, 06].

2.3.6. RDF (Resource Description Framework) :

RDF est un langage formel qui permet d'affirmer des relations entre des «ressources», il est utilisé pour annoter des documents.

Un document RDF est un ensemble de triplets de la forme <sujet, prédicat, objet> où < ressource, propriété, valeur>. Les éléments de ces triplets peuvent être des URIs, des littéraux ou des variables. Cet ensemble de triplets peut être représenté par un graphe (plus précisément un multiple-graphe orienté étiqueté), où les éléments apparaissant comme sujet ou objet sont les sommets, et chaque triplet est représenté par un arc dont l'origine est son sujet et la destination son objet.
Donc, il s'agit d'un formalisme utilisé pour représenter les propriétés d'une ressource et les valeurs de ces propriétés. Une ressource peut être une page HTML, un élément situé dans une page HTML, un ensemble de pages, bref tout objet qui peut être identifié de façon unique par un URI. RDF peut être vu comme un langage déclaratif de représentation de connaissances. Il traite uniquement de la représentation car il ne possède pas de capacités de raisonnement.

On va prendre un simple exemple que l'on veut chercher un livre dans bibliothèque. Ce livre est une ressource d'information. Pour le trouver, il faut d'avoir l'information concernée, par exemple : l'auteur, le nom du livre, la date de publicité ..etc. Autrement dit, ce sont les informations qui nous décrivent une information. Donc, ce type de cette information est appelé méta données qui sont principalement pour but de faire la recherche. Donc, RDF permet de représenter ces méta données attachées à des ressources [Abel, 2004][Ahcene , 05] .

Exemple RDF :

Figure 2.6 : Une illustration du RDF
Source : http://www.w3.org/TR/rdf-primer

La personne Eric Miller :

- Est identifier par l'URL http://www.w3.org/People/EM/contact #me
- Est de type (http://www.w3.org/1999/02/22-rdf-syntax-ns#type) Personne (http://www.w3.org/2000/10/swap/pim/contact#Person)
- a un nom complet (http://www.w3.org/2000/10/swap/pim/contact#fullName) Eric Miller

- a une adresse électronique
(http://www.w3.org/2000/10/swap/pim/contact#mailbox) em@w3.org
- et a un titre
(http://www.w3.org/2000/10/swap/pim/contact#PersonalTitle) Dr.

Les valeurs Eric Miller et Dr sont encadrées d'un rectangle, car contrairement aux autres éléments de ce graphe RDF, ce ne sont pas des URL, mais bien des littéraux. Un *littéral* est un objet qui n'est pas une URI mais bien un contenu réel, une valeur. Il peut être également typé, le type étant donné entre crochets derrière sa valeur.

2.3.7. RDFSchéma :

RDF ne propose pas d'outils permettant de spécifier le vocabulaire utilisé dans les descriptions. RDF Schema (RDFS [Brickley&Guha, 04][TA, 05]) est une extension de RDF permettant de décrire les concepts utilisés dans les descriptions et de définir des contraintes de type sur les objets et les valeurs des triplets. En fait, les triplets RDF sont les instances d'un RDFS.
RDFS fournit un schéma de base permettant une telle définition de vocabulaires dans un modèle objet: des classes de ressources et des types de propriétés.
RDF Schema est un vocabulaire permettant de décrire des vocabulaires simples.puisque, il est doté du nombre minimum de concepts nécessaires à la définition d'un vocabulaire.

• Il définit la notion de "classe" qui est un ensemble de plusieurs objets.
• Il définit la propriété particulière "est une sous-classe de" qui permet de définir qu'une classe est un sous-ensemble d'une autre classe.
• Il définit la classe des "ressources" qui est la classe mère de toutes choses :
- Tout est une ressource dans le web sémantique, sauf la notion de "littéral".
- Toute classe est une sous-classe de la classe des ressources.

• Il définit la notion de "littéral" qui est une valeur comme une chaîne de caractère ou des chiffres : ces choses ne sont pas des concepts et ne peuvent être manipulés comme tels.
• Il définit la propriété "s'applique à la classe" (range) permettant ainsi de spécifier le champ d'application d'une propriété.
• Il définit la propriété "est l'objet de la propriété" (domain) permettant ainsi de spécifier quelles sont les classes auxquelles ont peut affecter telle ou telle propriété.

2.3.8. OWL (Ontology Web Language):

Les langues plutôt ont été utilisées pour développer des outils et des ontologies mais elles n'ont pas été définies pour être compatibles avec l'architecture du WWW en général et le Web Sémantique en particulier. OWL en basant sur RDF nous donne les possibilités suivantes aux ontologies :

- Capacité d'être distribué à travers beaucoup de systèmes
- Compatibilité avec des normes du Web pour l'accessibilité et l'internationalisation
- Ouverture et extensibilité.

Jusqu'à maintenant, il y a pas mal d'organismes utilisant OWL avec les nombreux outils disponibles. Actuellement, la plupart des systèmes qui a utilisé DAML, OIL, ou DAML+OIL change maintenant à OWL. En outre, un certain nombre d'outils de langue d'ontologie, par exemple, Protégé qui est très forte et connu nous donne l'appui

pour OWL. De plus, il y a beaucoup d'ontologies disponibles sur le Web qui crée par OWL. Par exemple dans la bibliothèque de DAML [PHAN, 05] comme j'ai dit en haute, on peut utiliser les ontologies pour capturer la connaissance dans le domaine d'intérêt. Voilà, une ontologie va décrire les concepts dans ce domaine et les liens entre ceux. D'après [PHAN, 05], les différentes langues d'ontologie ont des avantages différentes. A ce moment, OWL est considéré par W3C comme une langue d'ontologie standard. Il a non seulement la capacité de décrire les concepts dans un domaine mais aussi d'un ensemble plus riche d'opérateurs, donc ces concepts bien définis et bien décrits. On peut construire des concepts complexes en basant les définitions des concepts plus simples. En outre, on peut vérifier si tous les rapports et les définitions dans l'ontologie sont conformés et identifier quels concepts s'adaptent sous quelles définitions. Donc, on peut maintenir la hiérarchie correctement entre les classes.
OWL peut faire un compromis entre son pouvoir expressif et son pouvoir de raisonnement parce qu'il fournit des sous langages de plus en plus expressifs conçu. Donc, OWL permet d'augmenter le sens du vocabulaire prédéfini dans une ontologie. On peut définir de chaque sous-langage grâce au son expression [PHAN, 05]. Par exemple: OWL-Lite est le moindre expressif. OWL-Full est le plus expressif. Et OWL-DL est entre celui du OWL-Lite et OWL-Full. On peut lui considérer comme une extension de OWL-Lite et OWL-Full est une extension de OWL-DL.

OWL Lite

OWLLite est le plus simple avec une structure syntactique. On l'utilise dans le cas une hiérarchie des classes simples et des contraintes simples. Par exemple, on l'envisage qu'il fournira un chemin rapide de migration pour les thésaurus existants et d'autres hiérarchies conceptuelles simples.

OWL DL

OWLDL est plus expressif que le OWL-Lite en basant sur des logiques de description [PHAN, 05]. Les logiques de description sont un fragment que l'on peut décider du FOL (First Order Logic [PHAN, 05]).
Donc, ils sont favorables au raisonnement automatisé. Il est donc possible de calculer automatiquement la classification hiérarchie et de vérifier les contradictions dans une ontologie. Pour choisir entre ceux, c'est basé sur les expressions simples avec OWL-Lite suffisantes ou non.

OWL Full

OWL-Full est le plus sous-langage expressif. On va l'utiliser dans le cas où on aurait besoin d'une expression très haute ou une grande possibilité de décision. Mais, on ne peut pas exécuter automatiquement un raisonnement dans OWL-Full ontologies. Pour choisir entre OWL-DL et OWL-Full, c'est dépend l'important de pouvoir effectuer le raisonnement automatisé ou de pouvoir employer la puissante de modélisation de façon plus facile comme méta-classes (classe des classes) par exemple. OWL ont trois syntaxes [PHAN, 05] : abstraire, basé RDF, et basé XML.

Par exemple: Avec OWL abstraire syntaxe, on va voir ce qui peut être nécessité une définition suivante:

Class(pp:old+lady complete intersectionOf (pp:elderly pp:female pp:person))
Class(pp:old+lady partial intersectionOf (restriction(pp:has_animal allValuesFrom (pp:cat))
restriction(pp:has_animal someValuesFrom(pp:animal))))

C'est-à-dire: chaque vieille dame doit avoir un chat d'animal (puisqu'elle doit avoir un certain animal et tout elle les animaux doivent être des chats.)

Quels sont les composants principaux d'une Ontologie OWL ?

OWL ontologies ont des composants pareils avec Protégé ontologie représenté sur Instances, Slots et Classes, mais avec les terminologies comme suivantes:

- **Individus:**

Il représente les objets dans le domaine. OWL n'utilise pas la supposition du nom unique comme Protégé. C'est-à-dire on peut référencer un même individu avec deux noms différents ou plus. Donc, il faut clairement expliquer que cet individu est identique avec un autre individu ou différente avec les autres.

- **Propriété:**

Elle est une relation binaire entre deux individus. Par exemple : la propriété *estCollegue* est un lien entre deux individus Tien et Hung..etc.. On a aussi les propriétés qui peuvent être inversé.
Par exemple: *superviserDe* est à l'inverse de *estSupervisePar*..etc. Elle peut avoir un single valeur, qui s'appelle fonction ou être transitif, symétrie [PHAN, 05] .etc.

- **Classes :**

Les OWL classes sont interprétées comme les ensembles avec des individus. Elles sont décrites en utilisant les descriptions formelles qui énoncent précisément les conditions pour être un membre de cette classe. Classes peuvent être organisés en hiérarchie de super classe et sous-classe, qui est également une taxonomie. Avec OWLDL, on peut faire automatiquement les relations entre deux classes grâce au raisonner. Parfois, le mot concept remplace une classe ou les classes sont des représentations concrètes des concepts.

2.4. Conclusion :

Cette partie a permis de présenter l'objet « ontologie » tel qu'il est connu et utilisé actuellement par la communauté représentant cette discipline.

Chapitre 3. Utilisation des ontologies pour la recherche d'information et l'extraction de connaissances:

3.1. Introduction :

Dans le contexte de la recherche d'information, un besoin existe de partager la signification de termes dans un domaine donné (parler le même langage), entre l'utilisateur est le contenu de la collection de documents. Or de nos jours, toute activité humaine spécialisée développe son propre jargon (langue de spécialité) sous la forme d'une terminologie et d'une conceptualisation associée spécifique.

L'existence de tels jargons entraîne des problèmes de compréhension et des difficultés à partager des connaissances entre les acteurs de l'entreprise, les services d'une entreprise et les entreprises d'une industrie, qui font des métiers différents.

Fondamentalement, le rôle des ontologies est d'améliorer la communication entre humains, mais aussi entre humain et ordinateurs et finalement entre ordinateurs.

Nous intéressons dans ce qui suit, à étudier le lien entre les ontologies et la recherche d'information.

On peut d'abord se demander si l'utilisation des ontologies dans la recherche d'information est un phénomène récent ou pas.

Nous utilisons le principe d'ontologie pour la recherche d'information dans notre vie quotidienne sans se rendre compte. Comme exemple élémentaire, ce que font les utilisateurs d'ordinateurs par exemple lorsqu'ils cherchent un fichier sur leur disque.

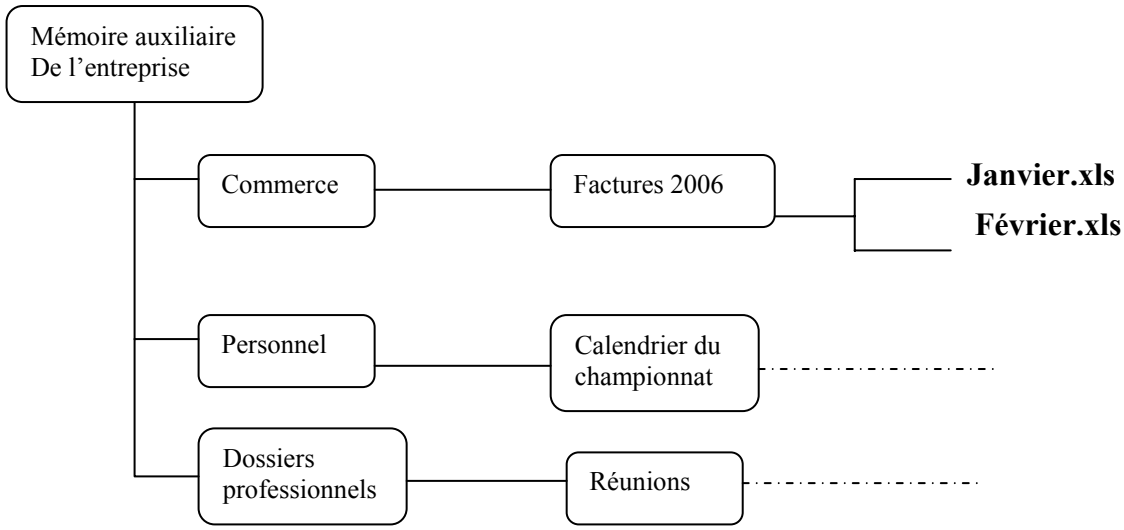


Figure 3.1 : Exemple de recherche de fichier dans une arborescence.

Pour accéder au fichier Excel de facture correspondant au moins de janvier de l'année 2006, l'utilisateur de l'ordinateur utilise des connaissances préalables pour distinguer le bon répertoire à chaque niveau de l'arborescente.

Dans le domaine de recherche d'information électronique tel qu'il est connu actuellement en utilisant des SRI, comment une ontologie peut-elle être associée au processus de recherche d'information ? Nous parlerons dans cette partie de l'utilisation des ontologies dans le domaine de la recherche d'information.

3.2. Principe d'utilisation des ontologies par un SRI :

L'utilisation d'ontologies dans un modèle de RI a pour finalité de spécifier des connaissances qui seront interprétables d'une part par l'utilisateur du système et d'autre part par le système lui-même. La connaissance qu'elle représente peut être utilisée à différents niveaux dans le processus de RI.

Les ontologies considérées doivent être adaptées aux tâches de RI visées et surtout apporter de la connaissance pertinente par rapport à l'information présente dans les corpus.

La figure suivante montre un exemple d'utilisation d'une ontologie de domaine dans le processus de recherche d'information entre la requête de l'utilisateur et le module chargé de la recherche effective des documents de base.

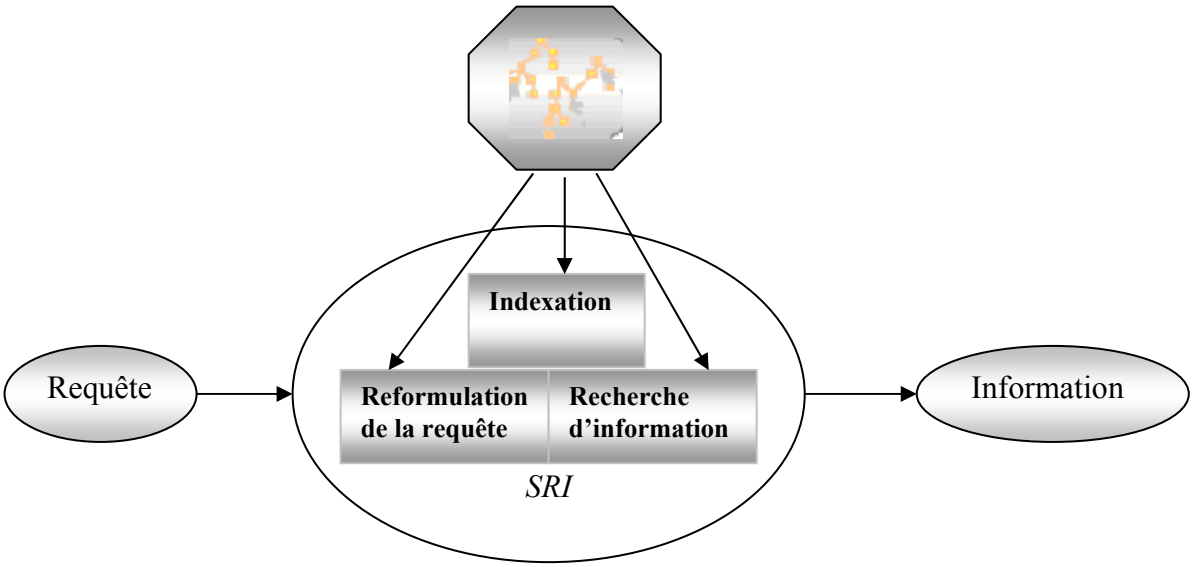


Figure 3.2 : L'ontologie greffée au processus de recherche d'information

Plus précisément pour un SRI, l'apport de l'ontologie peut être appréhendé sur trois niveaux comme schématisé en Figure :

- Au niveau du processus d'indexation des documents de la base : l'ontologie peut aider à l'indexation des documents, alors appelée indexation sémantique.

- Les ontologies peuvent également aider à la reformulation du besoin de l'utilisateur pour améliorer les requêtes utilisateurs, et à l'accès aux documents.

- Enfin l'ontologie peut être utilisée dans le modèle lui-même pour réaliser l'appariement entre le besoin et les documents.

En effet, de plus en plus de travaux en RI tentent d'améliorer l'indexation de textes ou la formulation de requêtes avec l'appui des ontologies. La recherche d'information (principalement sur le web), peut même être considérée comme un des champs d'applications favoris des ontologies. Par exemple, elles sont souvent présentées comme une pièce maîtresse dans le domaine du web sémantique.

3.3. Indexation sémantique: Indexation à partir d'ontologies

Nous distinguons deux types de démarches dans l'indexation sémantique : la démarche issue de la RI et la démarche issue du Web Sémantique.

La démarche issue du domaine de la RI consiste à choisir comme langage de représentation des documents, l'ensemble des concepts et instances de l'ontologie. L'utilisation d'ontologies sous forme de hiérarchies de concepts, ontologies légères ou lourdes est le prolongement de l'utilisation dans le cadre de la RI des ressources terminologiques décrites dans le chapitre 1 [Haav, 01][Nathalie, 05]. Les descripteurs ne sont plus choisis directement dans les documents ou dans un vocabulaire contrôlé (ou thésaurus) mais au sein même de l'ontologie. Les documents sont alors indexés par des concepts qui reflètent leur sens plutôt que par des mots bien souvent ambigus [Aussenac, 04][Nathalie, 05]. Il convient dans ce cas d'utiliser une ontologie reflétant le ou les domaines de connaissance abordés dans la collection documentaire. Il est en effet nécessaire de retrouver dans l'ontologie les concepts présents dans la collection pour indexer les documents à partir de toutes les thématiques abordées. Dans la littérature, il existe de nombreuses définitions de l'indexation sémantique. Certains auteurs différencient l'indexation sémantique de l'indexation conceptuelle [Mihalcea, 00][Nathalie, 05]. L'indexation conceptuelle repose, pour eux, sur des hiérarchies de concepts ou ontologies de domaine, alors que l'indexation sémantique repose sur l'utilisation d'ontologies génériques telles que WordNet. L'ontologie WordNet étant, selon nous, limitée par rapport à la sémantique qu'elle peut contenir, nous ne considérons pas que les mécanismes d'indexation qu'elle permet de mettre en place soient plus « orientés sémantique ». Les ontologies de domaine peuvent par leur formalisation représenter des ressources impliquant.

Nous entendons donc par indexation sémantique, l'indexation de documents à partir de n'importe quelle ontologie de domaine. L'indexation sémantique se fait en deux étapes. La première étape consiste à identifier les concepts ou instances de l'ontologie dans les documents. La deuxième étape pondère les concepts pour chaque document en fonction de la structure conceptuelle dont ils sont issus [Haav, 01][Nathalie, 05].

L'indexation sémantique est un type d'indexation qui s'inscrit également dans la démarche orientée Web Sémantique. Les précurseurs de cette nouvelle version du Web considèrent que les ressources participant au Web Sémantique seront toutes reliées entre elles par des relations sémantiques. Plus précisément, les données présentes sur le Web Sémantique seront modélisées sous forme d'ontologies où chaque ressource apparaît comme un élément de ces ontologies au même titre que la connaissance qui les décrit. L'objectif est donc d'ajouter au contenu du Web une structure formelle et de la sémantique (à travers des méta-données et de la connaissance) dans le but de permettre une meilleure gestion et un meilleur accès aux informations. Cette démarche repose sur des ontologies modélisant les objets du monde à travers les acteurs et entités que les documents constituent et comportent [Guha, 03][Nathalie, 05]. Elles peuvent être vues comme une représentation des méta-données explicitement ou implicitement présentes dans les documents. La phase d'indexation est aussi appelée annotation de documents. L'annotation de documents a pour but de représenter les informations relatives au média (date de création, taille, format d'encodage), les méta-données présentes dans les documents (auteurs, date de production), les index (les descripteurs du contenu du document), l'identifiant du document par le système (emplacement) et une vue sur le contenu (résumé ou extraits)

[Euzenat, 02] [Nathalie, 05]. La mise en place de cette nouvelle vision du Web dépend de la présence de ces méta-données. Un enjeu actuel du Web Sémantique est de définir des techniques permettant de les extraire [Guha, 03][Nathalie, 05]. La démarche orientée Web Sémantique a donc un double objectif : indexer le contenu des documents à partir des ressources permettant d'en extraire les concepts et instances mais aussi représenter les ressources en générant les méta-données correspondantes.

3.3.1. Identification des concepts et des instances existant dans l'ontologie :

La première étape de l'indexation conceptuelle consiste à identifier les concepts et/ou instances de l'ontologie apparaissant dans les granules.
Une première approche consiste à identifier ces éléments de l'ontologie manuellement dans les documents. Cette approche suivie dans [Vallet 2005] [Paralic 2003] [Kahan 2001][Nathalie, 05] est généralement réalisée par un expert et a pour intérêt d'être fiable car l'expert interprète la sémantique associée aux concepts dans l'ontologie et choisit le concept représentant au mieux la notion abordée dans le document. Cependant, même assisté par des traitements automatiques, ce procédé reste fastidieux, coûteux en temps et implique des erreurs [Erdmann 2000][Nathalie, 05].
D'autres approches visent à automatiser ce procédé. Cette démarche est légitime dans la mesure où l'utilisation d'une ontologie permet d'accéder à la connaissance et de la rendre manipulable par les systèmes. Dans ce cas là, les labels ou termes désignant les concepts ou instances sont recherchés dans les granules documentaires. Un concept (et une instance de concept) est en effet défini à partir d'un ou plusieurs labels représentant les variantes lexicales que peuvent prendre les termes définissant les concepts [Vallet 2005] [Kiryakov, 2004] [Guha 2003][Nathalie, 05].

3.3.1.1. Extraction des termes du document :

L'approche généralement suivie consiste à extraire des documents l'ensemble des termes y apparaissant et d'y rechercher les labels contenus dans l'ontologie. L'extraction de termes des documents se fait de la même façon que la recherche du langage de représentation classique. Les termes apparaissant dans un anti-dictionnaire peuvent être supprimés. Les expressions sont extraites soit statistiquement, soit syntaxiquement. L'extraction d'expressions est quasiment obligatoire car les labels des concepts sont souvent composés de ce type d'éléments.

3.3.1.2. Recherche des labels correspondant à des concepts de l'ontologie :

Les labels sont recherchés dans l'ensemble des termes extraits en favorisant la prise en compte des labels les plus longs et donc des concepts les plus spécifiques [Bloehdorn 2004] [Baziz, 05]. Par exemple, dans le cas où les labels « Madrid », « Real », et « Real Madrid » apparaîtraient dans le document, le label retenu - et donc le concept correspondant, - sera Real Madrid car l'expression formée de deux termes est plus précise que le ou les termes seuls. Plusieurs algorithmes ont été définis pour rechercher les labels les plus longs, ils consistent à faire varier la taille d'une fenêtre sur les mots de chacune des phrases des textes.

3.3.1.3. Désambiguïsation des labels :

Les labels peuvent cependant se rapporter à plusieurs concepts. Dans ce cas, un mécanisme de désambiguïsation du terme est mis en place afin d'identifier quel est le concept abordé dans le document. Il existe un grand nombre de techniques de désambiguïsation [Sanderson 2000][Nathalie, 05]. Les premières études faites sur l'intérêt d'utiliser la désambiguïsation en RI ont amené à des résultats variés, voire même contradictoires. Cependant, la conclusion qui peut être tirée de ces expériences est que des algorithmes de désambiguïsation de haute qualité sont nécessaires pour améliorer les performances du système [Sanderson 2000][Nathalie, 05]. Les techniques les plus simples considèrent les approches suivantes. La stratégie du « tout » correspond au cas dans lequel tous les concepts sont considérés. La stratégie du « premier » consiste à restituer le concept le plus fréquent dans le document ou bien dans la collection. La stratégie du « contexte » base la désambiguïsation sur la proximité sémantique des concepts candidats et du contexte dans lequel ils apparaissent dans les documents. Cette dernière variante peut être mise en place de diverses façons. Des règles syntaxiques et lexicales peuvent être générées manuellement, elles déterminent le sens d'un mot à partir des termes qui lui succèdent ou le précèdent dans son contexte. Cette approche a l'inconvénient de ne permettre la désambiguïsation que d'une faible proportion de termes. La désambiguïsation peut aussi s'appuyer sur des corpus déjà désambiguïsés. C'est le cas d'une des stratégies suivies dans [Mihalcea, 00][Nathalie, 05]. Le contexte du terme est représenté par les expressions qu'il forme à partir de tous les termes qui apparaissent directement après lui dans le corpus et directement avant lui. Le sens du mot est alors choisi à partir de son sens le plus courant dans les expressions représentant son contexte dans le corpus de référence SemCor [Miller, 93]. La limite de cette approche est que peu de ressources existent et qu'elles ne couvrent pas des domaines spécifiques. La désambiguïsation peut également reposer sur l'utilisation de ressources telles que des dictionnaires, des thésaurus ou des ontologies. L'utilisation de dictionnaires a pour principe de comparer les termes formant les différentes définitions du terme à désambiguïser avec les termes apparaissant dans le contexte du terme polysémique. Cette approche est suivie notamment dans [Lesk, 88] et [Mihalcea, 00]. Dans [Mihalcea, 00], le contexte d'un mot est représenté par les mots qui l'encadrent dans les documents dans une fenêtre de dix mots. Le sens choisi est celui dont la définition contient le plus de mots du contexte. Les relations entre termes (synonymies, est lié à) présentes dans les thésaurus et WordNet sont aussi utilisées pour désambiguïser les termes. Dans le cas où les ressources sont organisées hiérarchiquement, les mesures de similarités entre concepts telles que celles présentées dans la section 1 peuvent être utilisées [Banerjee, 02][Patwardhan, 03][Nathalie, 05].

3.3.1.4. Extraction de nouvelles instances :

L'extraction d'instances de concepts a pour but d'extraire les méta-données qui permettront de représenter les ressources dans le cadre du Web Sémantique. L'extraction d'instances repose sur des techniques du domaine de l'extraction d'information. De nombreuse plate-formes telles que Gate [Cunningham, 02] permettent de définir des patrons d'extraction ou d'utiliser des techniques reposant sur le traitement automatique des langues.
L'extraction d'instances de concepts peut se faire à partir de techniques d'extraction d'entités nommées, issues du domaine du traitement automatique des langues

[Kiryakov, 04][Nathalie, 05]. Une entité nommée est un nom ou syntagme nominal se rapportant à une entité comme, par exemple, une personne, une organisation ou une localisation. Les entités sont extraites à partir d'une base de connaissance qui, à partir de ressources lexicales, permet la détection automatique des entités. Les ressources lexicales décrivent par exemple les suffixes pouvant permettre la détection de noms d'entreprises ou de noms de familles ou de personnes. La base de connaissances contient un ensemble d'instances prédéfinies et décrites à partir d'axiomes. Un mécanisme d'inférence définit des règles permettant d'extraire de nouvelles instances. L'utilisation d'entités nommées et d'instances d'ontologie est une approche originale car les entités nommées sont rarement considérées en RI. La raison qui motive ces travaux est qu'une étude récente, faite sur les SRI, montre que 25% des requêtes contiennent des noms de personnes [Dumais, 03]. De plus, dans une approche traditionnelle par mot clé, l'utilisateur est obligé de spécifier à la fois le mot désignant l'instance qu'il recherche ainsi que les concepts auxquels se rapporte l'instance afin d'affiner sa recherche. Dans une approche conceptuelle, cette information n'a pas besoin d'être précisée car elle est connue par le système.
L'étape suivante consiste à pondérer les termes afin de mesurer leur représentativité du document.

3.3.2. Pondération des concepts et instances :

Le calcul du poids d'un concept ou d'une instance dans la représentation d'un granule peut être fait suivant plusieurs approches : statistiques ou conceptuelles.

3.3.2.1. Pondération statistique :

L'approche proposée dans [Vallet, 05] a pour but de calculer le poids des instances. Elle est inspirée de la méthode tf.idf.
Le poids $w_{i,j}$ d'une instance I_i dans un document D_j est calculé ainsi :

$$w_{i,j} = \frac{freq_{i,j}}{\max_k freq_{k,j}} * \log \frac{N}{n_i}$$

où $freq_{i,j}$ représente le nombre d'occurrences de I_i dans D_j, $\max_k freq_{k,j}$ est la fréquence de l'instance dans D_j, n_i est le nombre de documents annotés avec I_i et N est le nombre total de documents dans la collection.

Le nombre d'occurrences d'une instance a été défini comme le nombre de fois où le label de l'instance apparaît dans le texte, si ce document est annoté avec l'instance, ou 0 s'il ne l'est pas. Cependant, les résultats obtenus n'ont pas été satisfaisants car un grand nombre d'instances n'était pas reconnu à cause de lacunes à l'étape précédente correspondant à l'extraction des labels (non prise en compte des pronoms et périphrases notamment). Une approche similaire est présentée pour la pondération de concepts dans [Baziz, 05]. L'inconvénient de ces approches est qu'elles ne considèrent que les occurrences des concepts ou instances dans les documents et ne considèrent pas l'organisation conceptuelle dont ils sont issus. Une partie de la sémantique contenue dans les relations entre concepts est alors ignorée.

D'autres approches visent à combiner la pondération des concepts et/ou instances à partir de leurs occurrences dans les documents et leur place dans la représentation

conceptuelle. Elles reposent sur le calcul de similarité entre concepts présenté dans la section 1.

3.3.2.2. Pondération à partir de similarité conceptuelle :

Dans [Desmontils, 02] une approche est présentée pour indexer un ensemble de sites Web à partir d'une ontologie. Le pouvoir représentatif d'un concept prend en compte la fréquence d'apparition des termes désignant le concept dans les sites mais également ses relations avec les autres concepts du domaine. Plus un concept a des relations avec les autres concepts présents dans la page, plus il est représentatif de la page. Le pouvoir se calcule de la façon suivante :
Les termes d'une page Web sont tout d'abord extraits après analyse syntaxique (tree tagger) à partir de patrons (nom, nom+nom, nom+adjectif). Un premier poids, appelé poids de fréquence est calculé pour chaque terme en fonction de sa fréquence d'apparition et des balises html qui l'encadrent. Les coefficients correspondant à chaque balise sont attribués expérimentalement, par exemple, si un terme est encadré par la balise titre, le coefficient est 10, s'il est mis en gras, le coefficient est 2.
En supposant qu'un terme T_i apparaît p fois dans une page contenant n termes, $M_{i,j}$ étant le coefficient relatif à la balise encadrant l'occurrence j du terme T_i, le poids de fréquence P_freq de T_i est calculé ainsi :

$$P\_freq(T_i)=\frac{P(T_i)}{\max_{k=1..n}(P(T_k))} \quad \text{et } P(T_i)=\sum_{j=1}^{p}(Mi,j)$$

Ensuite, à partir de WordNet, l'ensemble des concepts relatifs à ces termes est généré sous forme de synset en prenant tous les sens définis. Un poids, appelé poids sémantique, est ensuite calculé en mesurant la similarité entre le concept donné et l'ensemble des autres concepts retrouvés.
Pour calculer la similarité entre 2 concepts, la formule sim définie dans [Wu & Palmer, 94][Nathalie, 05] est utilisée :

$$Sim(c1,c2)=\frac{2*depth(c)}{depth(c1)+depth(c2)}$$

où depth(cc) correspond au niveau de profondeur du concept cc dans la hiérarchie
et c est le concept subsumant c1 et c2.
Pour plus de détails sur la mesure de similarité voir la section 1.
Pour calculer le poids sémantique d'un concept dans une page, la somme des mesures de similarité du concept avec les autres concepts retrouvés de la page est calculée de la façon suivante :

$$P\_sem(synset_i(T_k))=\sum_{j\in[1,k-1]\cup[k+1,m]}\sum_{l=1}^{k}sim(synset_i(T_k),synset_l(T_j))$$

où $synset_i(T_m)$ représente le sens i dans WordNet retrouvé pour le terme T_m
Enfin, le pouvoir représentatif Rep du concept ou synset correspondant aux termes T_k est calculé en fonction de son poids sémantique et de son poids de fréquence :

$$Rep(synset(T_k))=\frac{\alpha*P\_freq(Tk)+\beta*P\_sem(symset(Tk))}{\alpha+\beta}$$

α et β sont fixés empiriquement à 1 et 2.

Les concepts retenus pour indexer chaque page sont ensuite choisis à partir d'un seuil sur ce pouvoir et de la présence de ce concept dans l'ontologie choisie pour indexer le corpus.

⏭Bilan :

En Bilan Dans le contexte de la RI, une ontologie représente la connaissance utile pour permettre une meilleure indexation. Il est donc indispensable qu'elle possède une forte composante lexicale afin de pouvoir mettre en correspondance les contenus des documents et les labels des concepts. Les mécanismes de pondération sont repris des mécanismes classiques de RI mais l'indexation est réalisée au niveau des concepts et non plus au niveau des termes.

Dans le contexte du Web Sémantique, les ontologies sont décrites formellement. Le principe suivi consiste à reposer sur une ontologie dans laquelle l'ensemble des concepts est défini et la phase d'indexation consiste à extraire des instances des concepts dans les documents.

3.3.3. Appariement à partir d'ontologies :

Les ontologies peuvent servir à calculer la similarité entre la représentation de la requête et la représentation des documents dans le cas où les deux représentations sont faites à partir des concepts d'une même ontologie.

Parmi les approches, l'approche est suivie dans [Andreasen, 03][Nathalie, 05] est que Les documents et requêtes sont représentés à partir du langage et de l'ontologie Ontolingua. Cette ontologie contient un ensemble de concepts et de relations entre concepts, dont la relation de subsomption. Elle est considérée comme un graphe orienté. L'avantage du calcul de la similarité est de classer les documents restitués par rapport à leur similarité à la requête, cette similarité reposant sur l'organisation des concepts dans l'ontologie. Le calcul de similarité s'appuie sur trois intuitions.

- La première intuition est que les documents liés au concept généralisant ou spécifiant le concept utilisé dans la requête peuvent intéresser l'utilisateur. Le calcul de la similarité prend donc en compte la distance séparant les deux concepts par la relation de subsomption. La similarité revient à prendre le nombre d'arcs séparant les deux concepts par le chemin le plus court à partir de la relation de subsomption.
- La deuxième intuition est que deux concepts ayant un concept les généralisant (ou subsumeur) commun sont plus similaires. Afin d'appliquer cette intuition, chaque concept est représenté par un ensemble flou à partir des concepts le généralisant. La similarité entre concepts est alors calculée à partir des éléments faisant partie de l'intersection entre les descriptions des concepts.
- La troisième intuition est que la similarité entre concepts doit prendre en compte les relations autres que les relations de subsomption. L'ensemble des concepts généralisant les deux concepts est alors considéré. Un sous-graphe de l'ontologie est construit à partir des concepts de cet ensemble pouvant être reliés dans l'ontologie par n'importe quel type de relation. La similarité est calculée par rapport aux nombres de noeuds ainsi connectés.

Cette approche est originale car elle calcule la similarité entre les concepts des documents et les concepts de la requête. La mesure de similarité proposée repose sur l'organisation des concepts dans l'ontologie. Cependant, aucune indication n'est donnée sur la combinaison des différents facteurs de la mesure. De plus, les auteurs ne considèrent pas le cas de figure suivant lequel plusieurs concepts sont retrouvés à la fois dans la requête et les documents et comment les différentes similarités sont combinées. Aucune évaluation n'est proposée.

3.3.4. Reformulation de requête à partir des termes de l'ontologie :

Il a été prouvé que la reformulation de requêtes a des effets positifs en RI . L'objectif de la reformulation est soit de limiter le silence (le silence fait référence aux documents pertinents mais qui ne sont pas retrouvés par le système) soit de réduire les risques de bruit (le bruit fait référence aux documents non pertinents retrouvés par le système). Dans le premier cas, la requête est étendue à partir de termes similaires à ceux de la requête initiale. Dans le second cas la requête initiale est étendue ou modifiée à partir de termes qui ajoutent de l'information complémentaire à la représentation du besoin. Il y a principalement deux approches permettant l'expansion de requêtes. La première consiste à utiliser des ressources, comme par exemple un dictionnaire, en étendant les requêtes à partir de nouveaux termes en relation avec les termes de la requête. La deuxième solution est la ré-injection de pertinence reposant sur l'analyse des termes contenus dans les documents jugés pertinents pour la requête initiale.

Un autre intérêt des ontologies est de permettre la désambiguïsation des termes de la requête. Dans [Guha, 03] la désambiguïsation se fait selon trois approches. La première consiste à choisir le concept dont les labels apparaissent le plus dans les documents. La seconde approche consiste à réaliser un profil utilisateur et à choisir le concept le plus proche de son profil. Finalement, la troisième prend en compte le contexte de la recherche et les documents recherchés par l'utilisateur jusque là aucune étude comparative n'est présentée.

Dans (Ka) [Benjamins, 99][Nathalie, 05], les pages Web sont annotées manuellement par des concepts d'une ontologie. Pour une requête donnée, tous les concepts liés aux termes de la requête sont inférés et ajoutés à la requête. Une interface a été développée pour assister l'utilisateur dans la formulation ou le raffinement de sa requête. Elle repose sur la visualisation de l'ontologie à partir de vues hyperboliques.

En conclusion, de plus en plus de travaux en RI tentent d'améliorer l'indexation de textes ou la formulation de requêtes avec l'appui des ontologies. La recherche d'information (principalement sur le web), peut même être considérée comme un des champs d'applications favoris des ontologies. Par exemple, elles sont souvent présentées comme une pièce maîtresse dans le domaine du web sémantique. Mais quelle est la contribution réelle des ontologies dans le processus de RI ? Sont-elles à la hauteur des espérances mises en elles ?

3.4. Apports de l'ontologie dans le domaine de la RI :

De manière générale, ce qui est attendu d'une ontologie, est qu'elle assure la réutilisation de connaissances. En recherche d'information, son apport est ciblé.
Nous donnerons dans ce qui suit d'après un rapport de [Masolo, 01] et le document [Malik, 02], une synthèse de la situation de ce domaine, du point de vue des ontologies utilisées dans les projets actuels. Pour rapporter sa conclusion, nous énumérons certains des gains visés attendus des ontologies :

- Les ontologies doivent réduire le silence dans les réponses aux requêtes :

Le but est de trouver autant de documents pertinents que possible dans une collection donnée.
A cet effet, les relations et les axiomes d'une ontologie, devraient fournir les moyens de rechercher quelques concepts qui ne sont pas explicitement écrits dans la requête : L'ontologie permet `a l'utilisateur lors d'une recherche sur le Web d'accéder non seulement aux documents liés aux mots clés de la requête, mais aussi à ceux qui sont liés ontologiquement `a ces derniers, ce qui rend la recherche encore plus pertinente. Elle a pour but de décrire des concepts et les relations qui les lient entre eux, et avec des règles de déduction les rendre plus compréhensibles et utilisables par les différents agents (humains ou logiciels). En dernier lieu, elle est interopérable.
Cette idée a été examinée très tôt dans des projets tels qu'OntoBroker et (KA)2 [Fensel, 01]. Des pages web ont été annotées avec des concepts à la main. Pour une requête donnée (éventuellement floue), tous les concepts possibles sont inférés, et toutes les pages annotées avec ces concepts sont recherchées. Dans Ontoseek, l'ontologie est représentée avec un graphe conceptuel et enregistré dans une base de données. Un document et une requête sont des sous graphes.

- Les ontologies doivent aider à réduire le nombre de réponses bruitées.

L'idée est d'ignorer les documents contenant les mots de la requête, mais avec un sens différent. Parce qu'elle contient des définitions non ambiguës, une ontologie devrait être assez précise pour fournir une définition unique à des termes, ou pour traiter la synonymie et l'ambiguïté d'une manière satisfaisante que ça soit dans la représentation d'ontologie ou dans l'annotation de documents. A cette fin, la conception de beaucoup d'ontologies pour la RI est contrôlée par une seule personne (comme dedans (KA)2 pour le domaine de l'ingénierie cognitive ou encore dans PICSEL quand une ontologie unifiée du tourisme est employée pour envoyer des questions à des sources diverses et hétérogènes). Dans le système Doc-Cube, un ensemble de termes synonymes (comme les Synsets dans Wordnet) est associé à chaque concept.

- Avec l'aide de l'ontologie, l'utilisateur peut exprimer son besoin plus facilement.
Afin de guider l'utilisateur, des étapes peuvent lui être suggérées pour préparer sa requête ou une nouvelle formulation avec des termes plus appropriés. L'ontologie permet d'établir une interface qui le guidera. Le parcourt de l'ontologie, mène à choisir des concepts et à définir une requête composée de concepts choisis et de leur description. Le fait d'exprimer une requête, revient alors à une instanciation des concepts de l'ontologie. L'interface de préparation de requêtes de PICSEL, propose ces fonctionnalités.

- Les ontologies doivent faciliter la recherche d'information dans des sources de données variées et hétérogènes et dans des domaines ouverts : pratiquement des travaux dans des domaines ouverts en général, conduisent à de petites avancées dans leurs résultats : Classer les documents prend du temps (principalement une fois fait à la main comme dans les ontologies SHOE ou OntoSeek) et peut mener à quelques erreurs, sauf si un système de désambiguïsation est disponible ; la réponse aux requêtes est beaucoup moins rapide qu'avec les systèmes traditionnels (surtout si plusieurs ontologies sont consultées comme dans Observer).
Cependant, la plupart des systèmes utilisent des ontologies spécifiques à des domaines fermés ("closed" environment). Le système recherche alors l'information dans un nombre limité de sources de données non évolutives. C'est le cas dans Planet-Onto ou pour un domaine donné, une ontologie centralisée fournie par un expert, est employée pour annoter de nouveaux documents dans le système d'information [Domingue, 99]. De nouveaux documents sont traités afin d'extraire de nouveaux termes spécifiques au domaine à ajouter à l'ontologie.
Pour résumer, nous dirons que très peu de projets concernent les aspects spécifiques de la recherche documentaire que nous venons de citer : grouper des documents (les relier aux classes caractérisées par une liste de mots clés ou de concepts) et leur récupération à l'intérieur d'un espace de l'information (l'espace étant constitué des documents qui partagent des critères sémantiques, utilisés pour la veille technologique par exemple). L'application des ontologies convient parfaitement au domaine de la recherche d'information. ceci est du au fait que :
- Le domaine est fermé, et couvre l'ensemble des documents dans la collection.
- Ces applications exigent juste une hiérarchie de concepts parce que l'expansion de requêtes ou leur spécialisation avec des types spécifiques de relations peut s'avérer dangereuse.

3.5. Les ontologies les plus connues :

Il est actuellement facile de recevoir des informations des organisations qui ont des ontologies sur le WWW. De nombreuses ontologies telles que les ontologies *Ontolingua* sur le serveur *Ontolingua* (1) (Farquhar *et al.* 1996) et *Wordnet*(2) (Miller 1990) à Princenton sont disponibles gratuitement sur la toile. D'autres ontologies, telles que les ontologies de *Cyc* (3) (Lenat *et al.* 1990) sont partiellement disponibles gratuitement sur le web. La majorité d'entre elles, cependant, ont été mise au point par des compagnies pour leur propre utilisation et ne sont donc pas disponibles. La *Ontology Page* (4) (également connue sous le nom de *Top*) et (*Onto*) *2Agent* (5) (Arpírez *et al.* 1998) (un moteur de recherche sur la toile s'appuyant sur une ontologie et qui aide à sélectionner des ontologies) peuvent aider à choisir des ontologies. Cette section introduit les ontologies les plus connues en prenant en compte la typologie d'ontologies énoncée ci avant.

3.5.1. Ontologies de représentation des connaissances :

Frame Ontology [Gruber, 93] :
L'exemple le plus représentatif des ontologies de représentation les connaissances est *la Frame Ontology* [Gruber, 93]. Elle saisit les primitives de représentation utilisées dans les langages de *frame*, telles que les classes, les attributs des sous-classes, les partitions de classes, les relations et les axiomes. Elle permet de codifier d'autres ontologies en ayant recours aux conventions habituelles des frames. Elle est

implémentée en *Kif 3.0* et constitue le matériau de construction de base des traducteurs d'*Ontology Server* [Rint, 99].

3.5.2. Ontologies de haut niveau :

Les ontologies de haut niveau fournissent des concepts généraux à partir desquels tous les termes des ontologies existantes peuvent être définis. Citons le treillis booléen de Sowa [Sowa, 97][Rint, 99] , le *Penman Upper Level* , *Cyc* (Lenat, 90), la proposition de très haut niveau de Guarino [Guarino, 97][Rint, 99], etc. En outre, des travaux sur une sorte d'ontologie «normalisée» de haut niveau ont été entamés au sein de l'Ansi en 1996.

L'*Ontologie méréologique* [Borst, 97][Rint, 99] :
Pourrait être l'exemple typique d'une méta-ontologie. Cette ontologie définit la relation *partie-de* et ses propriétés. Cette relation permet d'exprimer que des instruments sont formés de composants, qui peuvent eux-mêmes être constitués d'éléments plus petits.

L'ontologie CISC :
Une des premières ontologies développées dans un cadre informatique le fut dans le cadre du projet CYC du milieu des années 80. l'objectif était d'élaborer un système expert capable de comprendre et de parler un langage naturel, l'anglais. Le projet a consisté à constituer une ontologie en modélisant les connaissances de sens commun.
CYC est un projet qui a commencé en 1984 sous l'impulsion de D. Lenat. L'idée de Lenat est d'élaborer une base de connaissances dites « de sens commun », permettant à l'ordinateur de faire des inférences en apparence simples mais nécessitant en fait de nombreuses connaissances implicites [Lenat, 90][Lenat, 01][qa][Rint, 99]. Le nom CYC est emprunté au mot *encyclopédie* : Il s'agit de coder toutes les connaissances de type encyclopédique que les humains mettent en oeuvre pour comprendre un texte. Afin de comprendre la phrase *Napoléon est mort à Sainte-Hélène*, il faut savoir que Napoléon est un être humain, que les êtres humains sont mortels, etc. Lenat se fonde ainsi sur des phrases simples et sur les liens entre phrases pour essayer de déterminer tout ce qui est nécessaire à la compréhension.
Le projet initialement défini comprenait plusieurs étapes. De 1984 à 1990, environ 2 millions de connaissances de « sens communs » devaient être codées puis, progressivement au cours des années 1990, des mécanismes d'apprentissages à partir de dialogues ou de textes devaient être mis au point afin que le système puisse apprendre de nouveaux faits par lui-même à partir des années 2000. Un langage artificiel a été mis au point par l'équipe de CYC afin de dialoguer avec le système. Ce langage permet à CYC d'acquérir de manière autonome de nouvelles connaissances sur lesquelles peuvent s'appuyer de nouvelles inférences.
La base de connaissance de Cyc est une représentation formalisée d'une vaste quantité de connaissances humaines fondamentales: des faits, des principes de base, et des heuristiques pour le raisonnement au sujet des objets et des événements de la vie quotidienne.
La base de connaissances se compose des termes (qui constituent le vocabulaire de CycL) et d'assertions qui relient ces termes. Ces affirmations incluent à la fois les affirmations et les règles de base.
Cyc n'est pas un système basé sur les frames ; l'équipe de Cyc pense à une base de connaissances à la place, en tant que « mer d'affirmations », où chaque assertion n'est

pas plus « au sujet d' » un des termes impliqués qu'un autre. Elle consiste en un ensemble de termes et d'affirmations liées à ces termes. Elle se décompose, par ailleurs, en différentes « microthéories ». Chaque microthéorie rend compte seulement d'un point de vue important d'un domaine de connaissances. Certains domaines peuvent traiter plusieurs microthéories, qui représentent différentes perspectives et affirmations, divers niveaux de granularité et de distinction.
À l'heure actuelle, la base de connaissances Cyc contient des dizaines de milliers de termes et de nombreuses assertions saisies manuellement « à propos de » ou « impliquant » chaque terme. De nouvelles affirmations sont continuellement ajoutées à la base de connaissances par des humains.

3.5.3. Ontologies linguistiques :

D'après [Rint, 99], le Generalized Upper model [Bateman et al, 95], Wordnet [Miller, 90], et Sensus [Swartout et al, 97] représentent le mieux les ontologies linguistiques.

Generalized Upper Model :
L'ontologie est considérée comme une base d'un système multilingue de génération de textes.
GUM est une ontologie linguistique générale, indépendante de tout domaine et de tout type de tâche. Afin de pouvoir la transférer dans différentes langues, il a été prévu que l'ontologie Gum n'inclue que les notions linguistiques principales et leur organisation dans toutes les langues ; Elle omet ainsi les détails qui différencient les langues. Cette philosophie a permis d'utiliser Gum pour créer des ontologies pour des langues spécifiques, telles que l'anglais, l'allemand, l'espagnol et l'italien en rajoutant les traits sémantiques propres à chaque langue [Rint, 99].
L'ontologie GUM est un modèle d'organisation générale de concepts défini dans le langage de représentation des connaissances LOOM.

Wordnet :
Développé à l'université de Princeton par une équipe dirigée par Gearge Miller [Miller, 93] Sa naissance remonte à 1985. Le projet est amorcé par les mots du corpus Brown.
WordNet est une base de données lexicale pour l'anglais fondée sur des principes psycholinguistiques. Ce système de référence lexicologique en ligne est construit manuellement.
Les objets lexicologiques dans WordNet sont organisés sémantiquement (avec la distinction de base entre les noms, les verbes, les adjectifs, et les adverbes). Ses informations sont ventilées en unités appelées « synsets » en anglais, qui sont des jeux de synonymes interchangeables dans un contexte particulier utilisés pour représenter différents sens. Si un mot a plus d'un sens, il apparaîtra dans plus d'un synset.
WordNet contient une série de paires (w, m) où w est une série de caractères ASCII et m un élément d'un ensemble de sens, ou synset. Les synsets sont accompagnés dans leur plus grand nombre de glossaires explicatifs, et ils sont organisés en réseau sur la base de relations sémantiques, au nombre desquelles : l'antonymie, l'hyponymie, la métonymie, l'implication.
Néanmoins, beaucoup d'autres Instituts et groupes de recherche développent des WordNets semblables dans d'autres langues (européennes et noneuropéennes) utilisant le cahier des charges d'EuroWordNet. Si compatibles, ces WordNets peuvent

être ajoutés à la base de données et, par l'intermédiaire de l'Index, reliés à n'importe quel autre WordNet.
EuroWordNet est donc un ensemble de réseaux monolingues inspirés de WordNet et reliés entre eux. Il est crée dans une optique multilingue. Plusieurs objectifs sont liés à ce projet, tels que la construction de réseaux monolingues (qui sont des ontologies linguistiques qui effectuent des inférences) ; la recherche d'information translinguistique (grâce à l'augmentation des synonymes)

SENSUS :
L'ontologie SENSUS est un projet developé l'« University of Southern California » (USC). La construction initiale de SENSUS et de ses algorithmes d'alignement ont été exécutés par Kevin Knight et Steve Luk ; Les algorithmes postérieurs et la suite du travail sont effectuée par Eduard Hovy et Bruce Jakeway.
SENSUS est une ontologie basée sur le langage naturel qui a pour fonction de fournir une vaste structure conceptuelle aux travaux menés en matière de traduction automatique. Il a été mis au point en rassemblant et en extrayant des données de ressources électroniques telles que ; Penman Upper model, Ontos, WordNet (George Miller, Christiane Fellbaum; Université de Princeton) et des dictionnaires électroniques de langages naturels.
SENSUS est une taxinomie terminologique de 70000 noeuds, qui sert de cadre dans lequel on peut ajouter des connaissances supplémentaires.
Le premier objectif de SENSUS est la création et l'utilisation de grandes taxinomies de concepts (50000 ou plus) et d'ontologies pour le traitement du langage naturel en combinant les ressources en ligne telles que des dictionnaires et des thesaurus, des méthodes statistiques adaptées au texte, et des interfaces d'acquisition de connaissances humaines traditionnelles.
En particulier, créer et organiser une taxinomie de concepts de 70.000 éléments pour une utilisation dans PANGLOSS (traduction automatique), ou PENMAN (génération de phrase) et par la suite dans d'autres systèmes.
Cette recherche aborde le besoin d'acquérir de larges ressources en connaissances sémantiques et lexicologiques, à la fois pour le travail spécifique de PENMAN et pour permettre le partage de la connaissance entre les modules de PANGLOSS et d'autres sites.
L'ontologie est représentée en Loom, FrameKit, et Prolog.

3.5.4. Ontologies d'ingénierie :

Dans le domaine des ontologies d'ingénierie, les ontologies EngMath [Gruber et al , 94] et PhysSys [Borst, 97] méritent une attention particulière.

Ontologie *EngMath* :
EngMath est une ontologie *Ontolingua* mise au point pour la modélisation mathématique en ingénierie. Elle inclut des bases conceptuelles pour des grandeurs scalaires, vectorielles et tensorielles, des dimensions physiques, des unités de mesure, des fonctions sur les quantités et des quantités de dimensions.

Ontologie *PhysSys :*
PhysSys est une ontologie d'ingénierie destinée à modéliser, simuler et concevoir des systèmes physiques. Elle comporte trois ontologies d'ingénierie qui formalisent les

trois points de vue sur les outils physiques : présentation du système, comportement de processus physique et relations mathématiques descriptives.
Trois ontologies d'ingénierie formalisent chacun de ces trois points : une ontologie de composants, une ontologie de processus et l'ontologie *EngMath*. Les interdépendances entre ces ontologies sont formalisées comme des projections d'ontologies. Ces ontologies mettent en oeuvre d'autres méta-ontologies: méréologie, topologie et théories des systèmes.

5- d'après [Rint, 99], les ontologies qui représentent le mieux les ontologies dédiées à la modélisation d'entreprises sont l'*Enterprise Ontology* [Uschold *et a,* 96] et la *Tove Ontology* [Gruninger et al, 95][Rint, 99]. **L'*Enterprise Ontology :*** est un ensemble de termes et de définitions pertinent pour les entreprises commerciales et inclut des connaissances sur les activités et les processus, les organisations, les stratégies, le marketing, etc.

Tove Ontology :
Les ontologies élaborées dans le cadre du projet *Tove* (Toronto Virtual Enterprise) sont l'ontologie de conception d'entreprises, l'ontologie des projets, l'ontologie-agenda, ou encore l'ontologie des services.

6- L'ontologie *(KA)*[Benjamins *et a,* 99][Rint, 99] constitue un bon exemple d'ontologie dédiée à la gestion de connaissances, qui sera utilisée par le *Knowledge Annotation Initiative* de la communauté d'acquisition des connaissances. Cette ontologie servira de base pour annoter les documents sur internet de la communauté d'acquisition des connaissances de façon à fournir un accès intelligent à ces documents. Des spécialistes situés dans des zones géographiques différentes travaillent ensemble à la mise au point de cette ontologie.

3.6. Conclusion :

Ce chapitre a permis de présenter l'utilisation des ontologies pour la recherche d'information et l'extraction des connaissances tel qu'il est connu et utilisé actuellement.
Nous avons parlé de la relation entre les ontologies et la recherche d'information, où l'ontologie peut s'avérer comme un interlocuteur entre l'utilisateur et le SRI : l'utilisateur n'a pas accès aux données des sources, Il peut en ignorer le contenu, il dialogue avec le système dans le vocabulaire de l'ontologie.

Chapitre 4 :
Vers une approche basée Ontologie pour l'indexation automatique et la recherche d'information multilingue

4.1. Introduction :

Dans la première partie de ce mémoire, nous avons présenté un état de l'art des systèmes de recherche d'information. En particulier, nous avons essayé d'identifier leurs limites par rapport au traitement de l'information multilingue.
Nous avons présenté aussi, l'indexation suivant deux contextes de recherche d'information : la recherche à base de connaissances notamment les ontologies et la recherche multilingue.

Pour adresser le problème de langue les SRIM emploient comme base du procédé d'indexation les différentes approches connues dans la RIM tel que la traduction des documents, la traduction de requête. Vu les limites de ces méthodes, d'autres approches proposent d'adresser l'indexation multilingue en employant un langage pivot. Les documents et les requêtes étant traduits dans ce langage pivot.

Nous proposons dans cette partie une approche pour l'indexation et la recherche d'information pour un corpus trilingue : arabe, français et anglais. Le système proposé est fondé sur un formalisme de représentation de connaissances, plus précisément les graphes sémantiques [Roussey, 01] qui supportent une ontologie de domaine. Les documents et les requêtes sont aussi représentés dans ce formalisme.
L'ontologie du domaine constitue le noyau du système et est utilisée aussi bien pour l'indexation que pour la recherche. Le système d'indexation utilise une méthode d'extraction qui est basée sur le calcul de segments répétés en utilisant des filtres linguistiques. Quant au système de recherche, il est fondé sur la comparaison de graphes de requêtes et de graphes de documents.

4.2. L'extension du modèle des GC pour la RI:

Nous avons adopté dans notre approche les graphes sémantiques de Roussey comme langage d'indexation pour améliorer la description sémantique des documents dans un contexte multilingue. Ce formalisme est basé sur les graphes conceptuels (CG) de Sowa.

En effet, les graphes conceptuels se sont avérés plus appropriés dans le modèle de recherche documentaire.
Dans le formalisme de CG, une relation de spécialisation peut être calculée sur les CG par un opérateur de projection. Cette opération de projection permet de trouver les éléments spécifiques d'un graphe dans un autre graphe. Cette opération est employée dans des systèmes de recherche documentaire pour construire la correspondance entre le CG d'une requête et le CG d'un document. Si le CG d'une requête est projeté avec succès sur le CG d'un document, le document est retourné comme réponse à la requête.

La figure 4.1 est un exemple de projection du graphe H représentant « la comparaison de langues » sur le graphe G représentant « l'influence de la comparaison de langues occidentales sur la compréhension de ces langues ». Si dans la hiérarchie des types de concepts, le type 'langues occidentales' spécifie le type de concept 'langue' c'est-à-dire si *langues* est un terme générique de *langues occidentales* (ce qui se traduit dans T_C par *langues occidentales*≤ *langues*) alors, il existe une projection de H dans G (figure 4.1), car G contient un sous graphe dont tous les nœuds sont des spécifications des nœuds de H. Si H représente une requête et G index un document du corpus, le document référencé par G est jugé pertinent pour la requête.

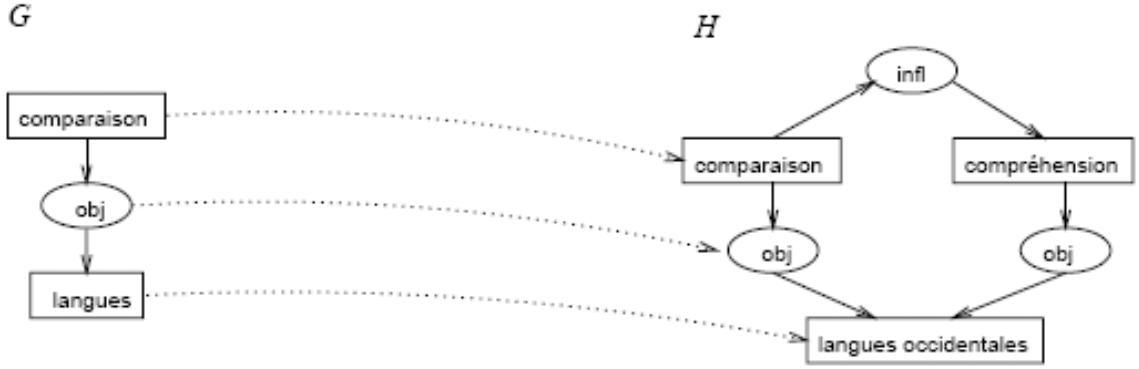


Figure 4.1- une projection

Ainsi, il est possible d'utiliser cette opération pour élaborer un mécanisme de recherche des documents pertinents étant donnée une base de documents D et une requête q :

$$\varepsilon = \{e_i \in D \,/\, \exists p \text{ projection de } q \text{ dans } e_i\}.$$

Néanmoins, l'opération de projection considère comme pertinents uniquement les index qui ont la même forme que la requête. Toutefois, l'utilisation de cette opération n'est pas satisfaisante car source de silence : un document qui ne serait qu'une réponse partielle à la question ne serait pas retrouvé. Plus précisément, pour qu'un document soit jugé pertinent pour une requête, le graphe conceptuel indexant le document doit contenir un sous graphe dont tous nœuds sont identiques ou spécialisent les nœuds du graphe représentant la requête. De plus l'opérateur de projection évalue l'existence d'une projection mais ne permet pas de classer les graphes en fonction de leur ressemblance. C'est à dire l'opération de projection renvoie un résultat binaire : il existe (au moins) une projection entre les deux graphes ou il n'existe pas de projection. Or, un système de recherche documentaire efficace ne peut fournir la liste des documents pertinents pour l'utilisateur mais plutôt une liste de documents classés dans l'ordre décroissant de leur pertinence estimée. L'utilisation de la projection ne permet pas d'obtenir un tel classement puisqu'elle donne un résultat binaire et il est donc nécessaire d'étudier d'autres opérations.
En effert, en recherche d'information, l'exactitude ou la précision n'est pas le critère unique d'un bon système de recherche d'information et s'oppose à un autre critère important, le rappel.
Les limites importantes de la projection ont conduit les chercheurs à proposer des extensions au modèle de CG, notamment, il s'agit de prolonger l'opération de projection afin de dépasser les limites ci-dessus.

[Genest, 00] propose un mécanisme pour que des documents proches d'une requête (par exemple un document générique par rapport à la requête) soient aussi jugés pertinents en se basant sur une extension de graphes conceptuels.
Tout d'abord, cette extension ajoute au support des graphes conceptuels de nouvelles relations de définition de types, autres que la relation de spécialisation. Par exemple, un type de concept n'est pas défini uniquement par rapport à ses ascendants (parents) ou descendants (enfants) dans la hiérarchie de spécialisation mais aussi par rapport à des types de concepts suivant une relation de composition ''partie de''. Ces nouvelles relations vont permettre de modifier les graphes pour découvrir de nouvelles projections. Ces modifications, sont appelées transformation.
La proposition de [Genest, 99] de principe de transformation pour les graphes conceptuels est définit comme suit :
'' Soient la représentation d'un document par un graphe conceptuel d, la représentation d'une requête par un graphe conceptuel q définis sur un même support S et un ensemble de règles de transformation K, la mesure de pertinence de d par rapport à q relativement à K est déterminée par la transformation minimale appliquée à d pour obtenir un d' tel qu'il existe une projection de q dans d''.

Les règles de transformation peuvent être comme suit :
– Changement du type d'un sommet concept, par l'utilisation du thesaurus : si un sommet concept est de type t, il peut être changé en un sommet de type t' tel que t et t' soient en relation dans le thesaurus;

- Changement du type d'un sommet relation ;
- Jointure entre deux sommets concepts (fusion voir exemple) ;
- Somme disjointe avec un graphe quelconque pour la recherche de réponses partielles (documents partiellement pertinents).
- Transformations plus spécifiques telles que la formulation de l'indexation, la manipulation de documents comprenant plusieurs parties et indexés par plusieurs graphes, etc.

Ce mécanisme, qui est une généralisation de l'opération de projection (si aucune règle de transformation n'est définie) permet de classer les documents par leur pertinence estimée. Plusieurs méthodes peuvent être utilisées pour cela, comme l'attribution d'une pondération à chaque type de transformation élémentaire. Dans ce cas, la pertinence d'un document peut être estimée par la somme pondérée des transformations élémentaires de d en d'.

Dans [Genest, 00], une série de transformation sur les graphes est définie, ainsi qu'un préordre sur les séquences de transformations ; les séquences de transformation sont donc ordonnées. Ce nouveau mécanisme de recherche détermine une séquence de transformations nécessaires à apporter aux graphes index pour qu'il existe une projection du graphe requête sur le graphe index. Les documents retrouvés sont donc classés en fonction de l'ordre des séquences de transformations à apporter à leur index pour qu'il existe une projection du graphe requête dans le graphe index.
Un exemple :
La figure 4.2 présente un exemple d'indexation de document et un graphe représentant une requête. L'indexation d peut être interprétée ainsi : « influence de la comparaison de langues européennes sur la compréhension de langues européennes », la requête q peut être interprétée par « influence de la comparaison de langues occidentales sur la compréhension de ces langues ». Le document semble donc

pertinent pour la requête considérée, alors que l'opération de projection ne permet pas d'arriver à ce résultat (il n'existe pas de projection de q sur d).

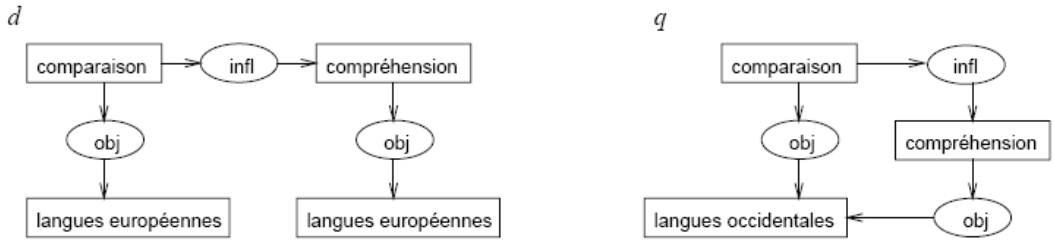


Figure 4.2 – Un graphe indexation d et un graphe requête q

Le mécanisme présenté ci-dessus estime pertinent le document représenté par d : La figure 4.3 présente une application des règles de transformation définies plus haut permettant d'aboutir à ce résultat (le sommet « entouré » dans chacun des graphes représente le résultat de l'application d'une règle de transformation par rapport au graphe précédent). Le graphe d_5 obtenu à partir de d par une séquence d'applications de règles est tel qu'il existe une projection de q dans d_5. Le document représenté par d est donc pertinent, et sa pertinence est estimée par les transformations de d en d_1, d_1 en d_3 et d_3 en d_5.

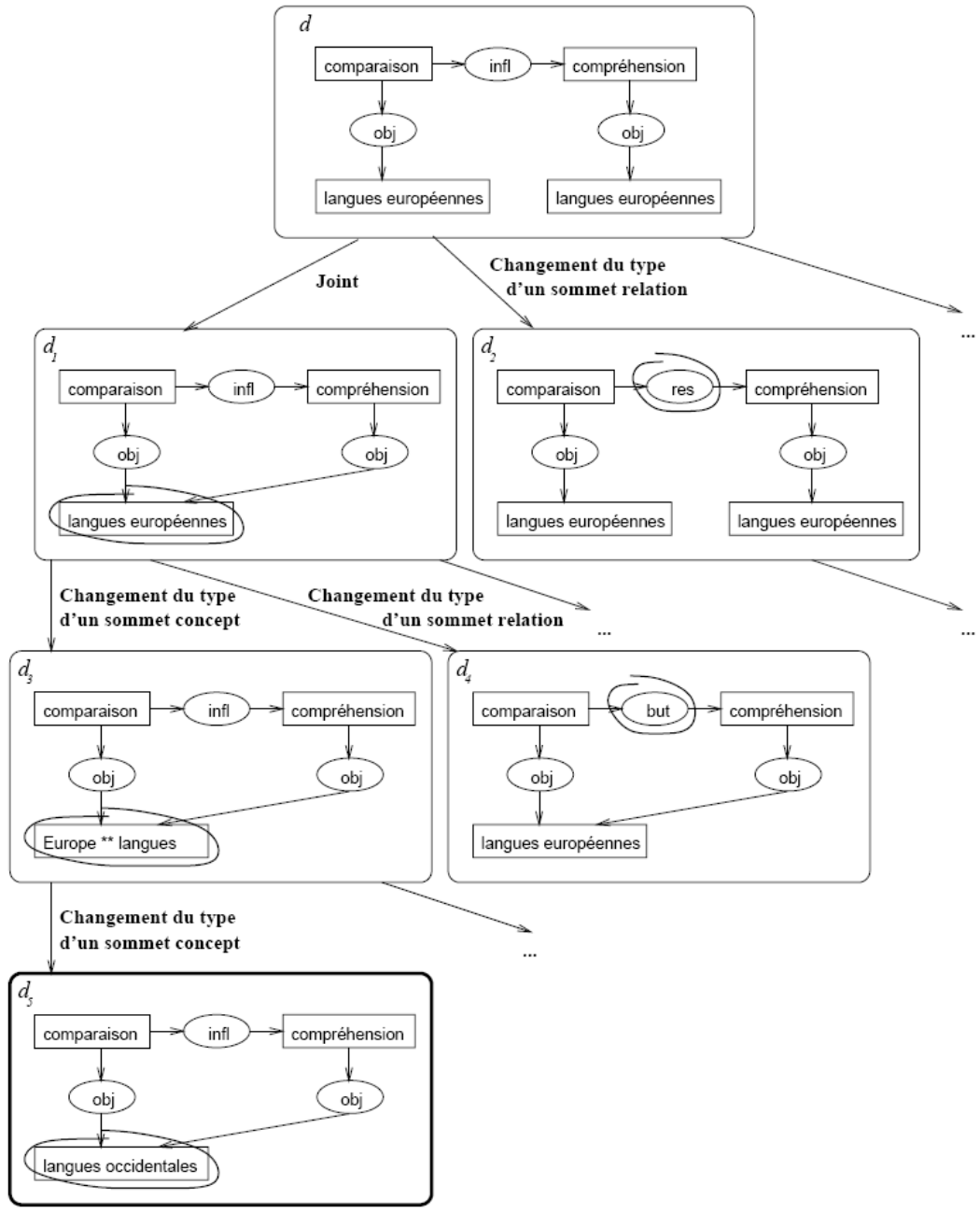


Figure 4.3 – Application de règles de transformations

Nous verrons dans le paragraphe suivant que [Roussey, 01] propose une simplification de ce modèle générique pour prendre en compte **l'aspect multilingue** en ajoutant au support des graphes conceptuels la notion de **vocabulaire** et au graphe la notion de **label**.

4.3. Formalisme des graphes sémantiques :

Comme nous l'avons dit plus haut, nous adoptons les graphes sémantiques (GS) de Roussey [Roussey, 01] pour améliorer la description sémantique des documents dans un contexte multilingue. Ce formalisme est basé sur les graphes conceptuels de [Sowa, 84]. Les graphes sémantiques permettent de représenter formellement,

rigoureusement, synthétiquement et élégamment une structure sémantique quelconque : mot, texte entier, personnage, action, etc. [Lounis, 06].
Un graphe sémantique est un ensemble de sommets concepts reliés entre eux ou non. La formalisation a pour but de mettre l'accent sur les couples de concepts reliés par une relation. Un arc est alors défini comme un couple de concepts étiqueté par un type de relation.
Les graphes sémantiques proposent d'améliorer la sémantique des documents dans le corpus multilingue par l'instanciation du modèle de CG. Afin de construire un formalisme qui sera plus près des thesaurus comme des langues. Ce modèle représente la première tentative pour créer un language de représentation basé sur le graphe de document multilingue.
Formellement, un graphe sémantique $Gs= (C, A, \mu, labelC, \nu, labelR)$ défini sur un thesaurus sémantique *M*, est un multigraphe, non nécessairement connexe, où :
- *C* est l'ensemble des sommets concepts de *Gs*.
- $A \subset C \cdot C$ est l'ensemble des arcs de *Gs*. On notra un arc a comme équivalent au couple de concepts (c,c') $\in$ C x C : par exemple, a = (c, c') $\in$ A.
- $\mu\ C \rightarrow T_C$ est une application qui à tout concept, $c \in C$ associe une étiquette $\mu\ (c) \in T_C$, $\mu\ (c)$ est appelé le *type* de *c*.
- *labelC* est un ensemble d'applications $labelC = \{ labelC^{VL1}, \ldots, labelC^{VLj}, \ldots, labelC^{VLp}\}$ telles que $labelC^{VLj}: C \rightarrow V_{Lj}$, qui à tout concept $c \in C$, associe une étiquette correspondant à un terme d'une langue donnée *Lj*, $labelC^{VLj}\ (c) \in V_{Lj}$. $labelC^{VLj}\ (c)$ est appelé le *label* de *c* pour la langue *Lj*. On note *labelC (c)* le *label* de *c* lorsque aucune indication sur la langue n'est donnée.
- $\nu : A \rightarrow T_R$ est une application qui à tout arc $a \in A$, associe une étiquette $\nu\ (a) \in T_R$, de l'ensemble des relations. $\nu\ (a)$est appelé le *type* de a.
- *labelR* est un ensemble d'applications $labelR = \{ labelR^{VL1}, \ldots, labelR^{VLj}, \ldots, labelR^{VLp}\}$ telles que $labelR^{VLj}: A \rightarrow V_{Lj}$, qui à tout arc, $a \in A$ associe une étiquette correspondant à un terme d'une langue donnée *Lj*. $labelR\ (a)^{VLj} \in V_{Lj}$. $labelR^{VLj}\ (a)$ est appelé le *label* de *a* pour la langue *Lj*. On note *labelR (a)* le *label* de *a* sans indication sur la langue donnée.

Les graphes sémantiques doivent répondre à un certain nombre de contraintes.
1. A obéit aux contraintes fixées par l'application σ définie dans le thésaurus sémantique. Pour tout $a = (c, c') \in A$, les contraintes sur la signature du type de a, $\nu\ (a) = r \in T_R$., doivent être vérifiées par les concepts de a. Autrement dit $\mu\ (c) \leq_C \sigma_1(r)$ et $\mu\ (c') \leq_C \sigma_2(r)$. En prenant l'exemple du type de relation T_{r3} (est dans), T_{r3} a comme signature $\sigma(T_{r3}) =$ (universe, lieu). D'après la contrainte que nous venons d'énoncer, toute relation typée par ''est dans'', devra avoir comme premier argument un type de concept spécialisé par *universe* et comme second argument un type de concept spécialisé par lieu.
2. $labelC^{VLj}$ est la conjonction des deux applications μ et λ_C^{VLj}. En effet, pour tout $c \in C$, $labelC^{VLj}(c) = \lambda_C^{VLj}(\mu(c))$.
3. $labelR^{VLj}$ est la conjonction des deux applications ν et λ_R^{VLj}. En effet, pour tout $a \in A$, $labelR^{VLj}\ (a) = \lambda_R^{VLj}(\nu\ (a))$.

D'après les définitions, il existe plusieurs représentations des graphes sémantiques. En effet, des étiquettes différentes sont proposées pour chaque composant du graphe.

1- La première représentation possible consiste à étiqueter les sommets concepts et les arcs d'un graphe sémantique par leur type. Donc comme nous le montre la Figure 4.4, pour un sommet *c*, son étiquette est $\mu(c)=t_{c3}$. Cette représentation des graphes sémantiques est Utilisée pour stocker les index de notre système documentaire multilingue.

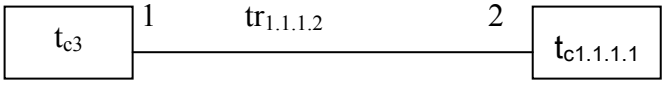


Figure 4.4 : Un exemple de graphe sémantique étiqueté par les types

2- La seconde manière de représenter un graphe sémantique consiste à étiqueter tous ses sommets et ses arcs par leur label dans une langue donnée. Si la langue fixée est l'anglais, le vocabulaire employé est donc le vocabulaire anglais *Veng*. Pour un sommet *c,* son étiquette est $labelC^{Veng}(c)= \lambda_C{}^{Veng}(\mu(c)) = \lambda_C{}^{Veng}(t_{c3})$= Fuel. Il existe autant de représentations d'un graphe sémantique sous cette forme, qu'il existe de vocabulaires disponibles dans le thesaurus sémantique. Par exemple, la Figure 4.5 présente deux représentations du même graphe sémantique. Dans la première, les étiquettes sont les labels français du graphe sémantique et dans la seconde, les labels sont anglais.

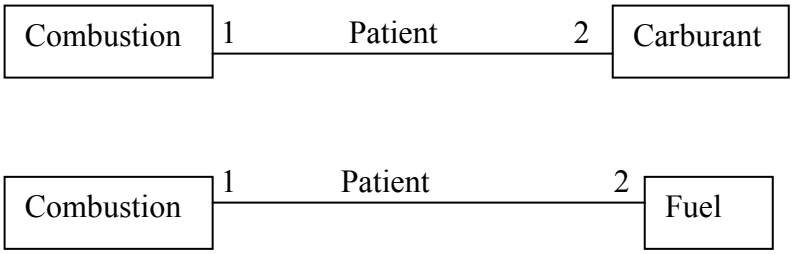


Figure 4.5 : Des exemples de graphes sémantiques étiquetés par des labels

En résumer, pour ce graphe sémantique composé d'un seul arc a = (c,c') :

C = {c, c'}

A = {a}

μ : $\mu(c) = t_{c3}$ $\mu(c) = t_{c1.1.1.1}$

ν : $\nu(a) = tr_{1.1.1.2}$

$labelC = \{ labelC^{Vfr}, labelC^{Veng} \}$

$labelC^{Vfr}$: $labelC^{Vfr}(c) = \lambda_C{}^{Vfr}(\mu(c)) = \lambda_C{}^{Vfr}(t_{c3})$= 'Combustion'

$labelC^{Vfr}(c') = \lambda_C{}^{Vfr}(\mu(c')) = \lambda_C{}^{Vfr}(t_{c1.1.1.1})$= 'Carburant '

$labelC^{Veng}(c)= \lambda_C{}^{Veng}(\mu(c)) = \lambda_C{}^{Veng}(t_{c3})$= 'Combustion'

$labelC^{Veng}(c')= \lambda_C{}^{Veng}(\mu(c')) = \lambda_C{}^{Veng}(t_{c1.1.1.1})$= 'Fuel'

$labelR = \{ labelR^{Vfr}, labelR^{Veng} \}$

$labelR^{Vfr}$: $labelR^{Vfr}(a) = \lambda R^{Vfr}(\nu(a)) = \lambda R^{Vfr}(tr_{1.1.1.2})$= ' Patient'

$labelR^{Veng}$: $labelR^{Veng}(a) = \lambda R^{Veng}(\nu(a)) = \lambda R^{Veng}(tr_{1.1.1.2})$= ' Patient'

4.4. Observations sur les langages documentaires :

Comme nous avons cité dans le chapitre1, lorsqu'un document est indexé, son contenu est décrit par un ensemble de terme, appartenant à un langage documentaire, qui vont constituer l'index du document. Ces descripteurs d'un document représentent

des notions ou concepts qui sont le résultat d'une abstraction. Nous reprenons la définition de notion donnée dans [FEL, 84][Roussey, 01] :'' les notions sont les représentations mentales des objets individuels. Une notion peut ne représenter qu'un seul objet individuel ou, par abstraction, comprendre tous les individus qui ont en commun certain caractère. Elle sert de moyen d'agencement mental (classification) à l'aide de symbole linguistique (terme, lettre, symbole graphique), de moyen de communication''. Pour simplifier nos explications, nous emploierons le mot notion comme synonyme de concept générique c'est à dire une abstraction faite sur un ensemble d'objets ou d'individus.

L'ensemble des termes reconnus par le SRI constitue le langage d'indexation.

C'est à dire pour la représentation du continu des documents, les systèmes de recherche documentaire utilisent un langage, appelé *langage documentaire*, qui définit le vocabulaire et les règles qui doivent être respectées pour la création d'indexations. Les contraintes imposées par l'utilisation de ce langage ont pour but de permettre la définition d'un mécanisme de recherche plus efficace que si l'indexation était totalement libre. Ainsi, la plupart des langages documentaires imposent le vocabulaire devant être utilisé, permettant ainsi d'ignorer, dans les indexations, les problèmes d'homonymie et de polysémie courants dans la langue naturelle. Par exemple, le terme *français* peut désigner une personne (un Français) ou une langue (le français). La définition d'un *vocabulaire contrôlé* peut interdire l'utilisation d'un tel terme au sens ambigu et imposer l'emploi de *français (langue)* ou *français (peuple)* afin qu'il n'y ait pas d'ambiguïté sur les concepts représentés par les indexations.

Les langages documentaires ont pour but de répertorier les notions d'un domaine et non de répertorier toutes les relations, individus, objets d'un domaine. C'est pourquoi un thésaurus, répertoriant les termes d'un langage documentaire, n'est pas un état des lieux des objets existants d'un domaine, mais un répertoire des notions existantes dans un domaine. Par exemple, un thésaurus de l'automobile ne va pas lister toutes les voitures d'une marque. Comme le point commun entre toutes ces voitures est leur marque, cette marque pourra faire partie d'un thésaurus de l'automobile.

Pour améliorer les possibilités de description, les relations ont pu être introduites dans les langages documentaires mais elles ont directement contribué à augmenter le nombre d'index possibles pour un même document. Cet effet appelé "paraphrase" diminue le nombre de documents pertinents retrouvés par le SRI c'est-à-dire diminue les performances d'un système de recherche.

Suivant cette constatation, Roussey [Roussey, 01] propose de transformer le formalisme des graphes conceptuels afin qu'il soit plus proche des langages documentaires.

En particulier, il s'agit de limiter les concepts à la notion de concepts génériques, c'est-à-dire sans marqueur. En effet, les marqueurs sont utilisés pour identifier un objet du domaine. Dans notre cas, le contenu des documents est représenté par des notions et non par des individus ou des objets particuliers. C'est pourquoi Roussey propose de simplifier le formalisme des graphes conceptuels en éliminant les marqueurs, et mettre l'accent sur une nouvelle notion : les vocabulaires. Un vocabulaire défini pour une langue donnée les labels des notions génériques décrites dans le support (les termes représentent ces notions). Nous intitulons le support et son ensemble de vocabulaires un thésaurus sémantique. De plus, nous ne travaillerons que sur la forme normale d'un graphe : un concept ne pourra apparaître qu'une seule fois dans le graphe, ce qui limite le nombre d'index possibles pour un même document.

Nous nous intéressons à trois langues arabe, français et anglais donc nous avons 3 vocabulaires, un par langue [Aliane et al, 06].

4.5. Un modèle de SRIM basé sur l'ontologie du domaine :

4.5.1. Vue globale du modèle :

Notre approche est centrée autour d'une ontologie du domaine qui est fondée sur le formalisme de graphes sémantiques.
La première étape de notre système est une construction manuelle de l'ontologie (Thésaurus sémantique). La caractéristique du système consiste en une conception basée sur l'interaction homme-machine. En effet, le rôle de l'expert humain est primordial pour atteindre les objectifs de performance du système et cela en tant que gestionnaire de ressources de connaissances qui peuvent varier en « intelligence » ou en « puissance ». Dans cette étape, des interfaces hautement visuelles sont offertes pour l'expert humain en vue de la construction de son thésaurus sémantique. Le système est indépendant du domaine dans le sens où différents experts peuvent créer différentes bases de connaissances. Donc, la conceptualisation du domaine ne pourra être modifiée que par les acteurs de l'indexation, et les vocabulaires évoluent en fonction de la terminologie des documents du corpus.

L'architecture globale de notre système se présente comme suit :

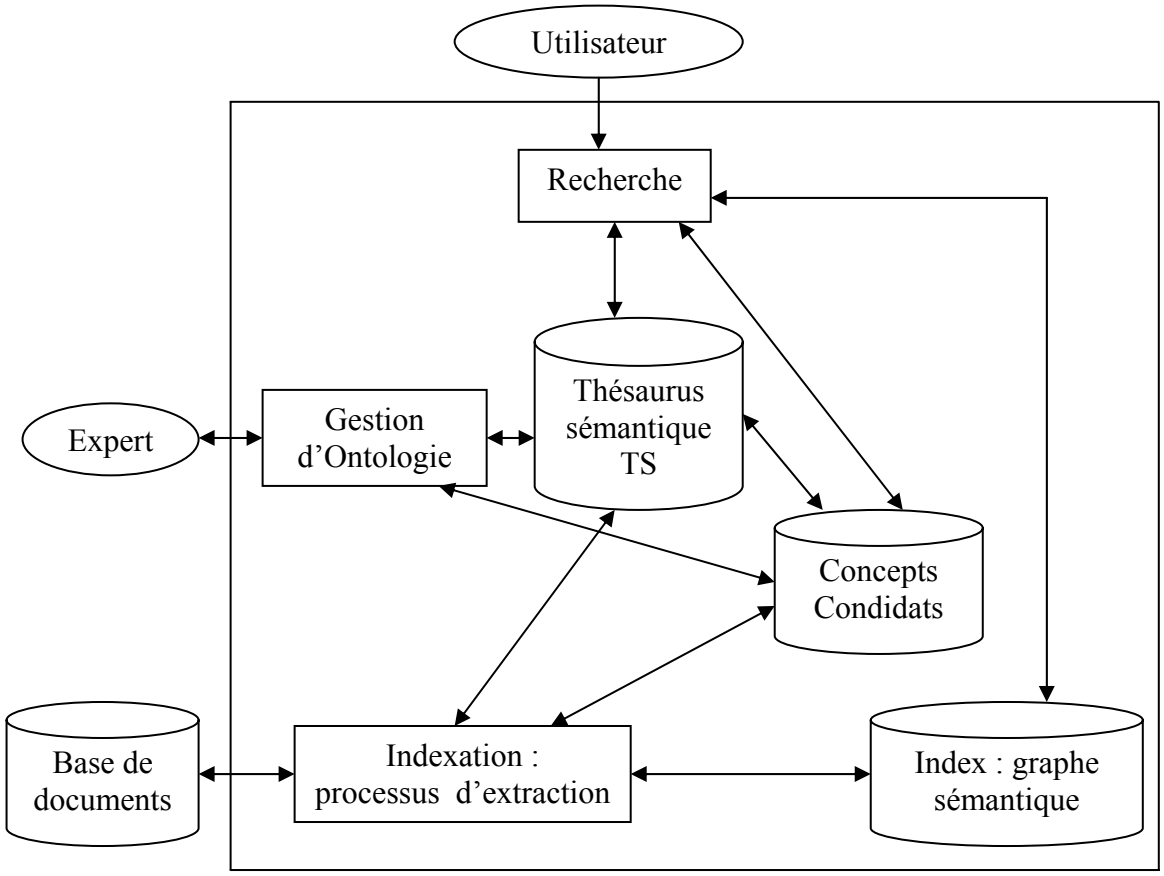


Figure 4.6 : Architecture d'un système de RIM basée sur Thésaurus sémantique

4.5.2. L'ontologie du domaine : Thésaurus sémantique :

Pour prendre en charge le problème de la langue, nous avons vu que les SRI utilisent un langage pivot comme base d'indexation. Aujourd'hui, les ontologies connaissent un réel succès dans la communauté des chercheurs en RI. C'est pourquoi nous avons choisi de construire notre langage-pivot comme ontologie de domaine fondée sur le formalisme des graphes sémantiques.

L'ontologie (thesaurus sémantique) qui représente le noyau de notre système, différencie deux niveaux de connaissance :

1- Le niveau conceptuel :

Modélise le domaine d'étude formé de concepts et de relations conceptuelles. Dans notre cas, il s'agit d'une conceptualisation du domaine ou support résultant d'un consensus entre les différents acteurs de l'indexation dans un domaine particulier. Cette conceptualisation équivaut au **support** du modèle des graphes conceptuels de Sowa [SOWA, 84].

2- Le niveau terminologique :

Est composé de l'ensemble des termes. Le terme est défini comme la manifestation linguistique d'un concept repéré dans un texte, dont il peut être considéré comme le **label**. "Le terme est donc un signe linguistique qui se distingue du mot de la langue par sa fonction de référence à une notion du domaine" [SEGU, 1997][Roussey, 01]. L'ensemble des termes d'une langue L_J constitue un vocabulaire du domaine soit V_{Lj}.

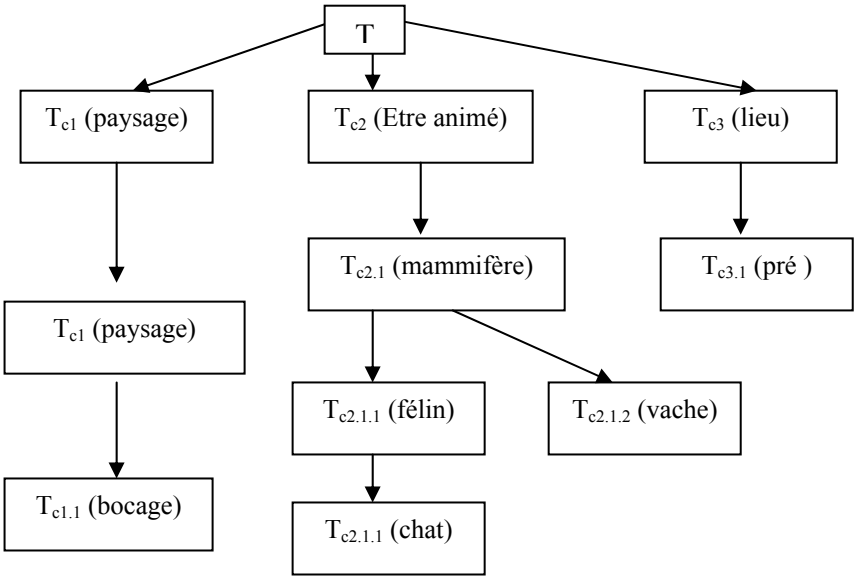
Le thesaurus sémantique se compose donc d'une conceptualisation du domaine *S* et d'un ensemble de vocabulaires *V*. La conceptualisation formelle *S* est lisible par un umain en remplaçant ses entités formelles de S par des termes du vocabulaire associés à la langue considérée. Cette conceptualisation pourra donc être présentée à l'utilisateur dans le vocabulaire de la langue de son choix.

4.5.2.1. La conceptualisation du domaine ou Support :

Un support comprend deux hiérarchies de types de concepts et de relations, les relations pouvant être munies de signatures. Les deux hiérarchies sur les ensembles de types de concepts et de relations sont interprétées comme des liens *Sorte de* qui forment deux hiérarchies partiellement ordonnées notées '≤'. Donc, le vocabulaire utilisé dans les graphes sémantiques est définit par un support.
Les types de concepts représentent des points de vue sur les objets du domaine. Par exemple, la notion de véhicule peut être vue sous plusieurs point de vue : il s'agit soit du moyen de locomotion soit de la machine utilisant un moyen de propulsion. Donc, dans le support, chacun de ces points de vue sera représenté par deux types différents.
Les types de relations définissent les liens possibles entre les notions.
Dans le support, les types sont repérés par des clés numériques et non par des termes, pour deux raisons : éviter toute interprétation a priori et ne pas figer l'appartenance d'un type à une langue particulière.

Plus formellement, la définition du support est la suivante :
Un support S est un triplet S = (TC, TR,σ) tel que :
- TC, est l'ensemble des types de concepts. Il est partiellement ordonné par la relation de spécialisation/généralisation, notée ≤ et il admet un plus grand élément, noté T appelé type universel.
Idc1.1≤idc1 signifie que le type idc1.1 est plus spécifique que le type idc1. La relation inverse a noté le ≥ est la relation de généralisation.
TR, est l'ensemble des types de relations. Il est partitionné en ensembles de types de relations de même arité : TR = TR1 ∪ TR2 ∪ …∪ TRj ∪ … ∪ TRn, où TRj est l'ensemble des types de relation d'arité *j* (*j* 0). Tout T_{Rj} est ordonnée par une relation de spécialisation/généralisation, notée ≤ et il admet un plus grand élément, noté Tj. Dans notre cas est noté T_2 (notre travail est lié au traitement de la langue naturelle, les type de relation correspondent au cas des verbes où les relations sont binaires. Donc, nous allons limiter aux types de relation d'arité 2).
- σ, appelée la signature, associe à tout type de relation, le type de concept le plus générique utilisable comme argument de la relation. C'est plus précisément une application qui, à tout $t_r \in T_{Rj}$, associe un j-uplet $\sigma(t_r) \in (T_C)^j$, et vérifie que, pour tout élément de $t_{r1}, t_{r2} \in T_{Rj}$, si $t_{r1} \leq t_{r2}$ alors $\sigma(t_{r1}) \leq \sigma(t_{r2})$. L'ordre considéré sur les signatures est l'ordre produit sur $(T_C)^j$. Autrement dit, le type associé au $k_{ième}$ argument de t_{r1} est plus spécifique que le type associé au $k_{ième}$ argument de t_{r2}. On notera $\sigma_i(t_r)$ le $i_{ème}$ argument de $\sigma(t_r)$. La Figure 4.7 présente deux exemples de signatures vérifiant les conditions précédentes.

Tc : ensembles des types de concepts.



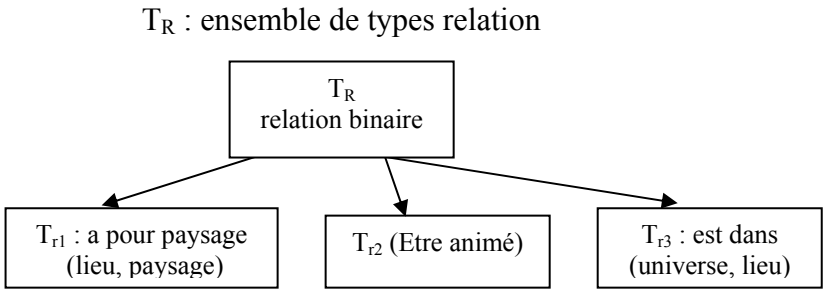


Figure 4.7 : un exemple d'ensembles de types de concepts et de types de relations ordonnés par les relations de spécialisation.

4.5.2.2. Hiérarchie des types de concepts:

La hiérarchie des types de concepts correspond à une modélisation conceptuelle d'un domaine donné, cette modélisation représente donc un point de vue sur le domaine et avec précision, il représente une vue d'indexation du domaine.
Le processus de la conceptualisation est le suivant :
- identifier le domaine ou les documents (corpus) pour les classer (pour indexation). C'est à dire délimiter le domaine en identifiant quel genre de documents seront indexés à l'aide de cette modélisation. Comme l'énonce [SEGU, 97][Roussey, 01], le corpus délimite le domaine.
- Identifier les concepts ou la connaissance de domaine. C'est à dire identifier les notions du domaine, en déterminant ce qui peut intéresser les lecteurs dans les documents.
- Normaliser ces concepts (notions) sous forme des types de concepts. C'est à dire leur associer un identifiant, que ne soit pas un terme, car plusieurs termes peuvent la désigner, puis leur associer une définition, même si la description d'une notion est difficilement exhaustive. ''Normaliser une notion c'est fixer une référence de définition et d'interprétation'' [SEGU,97][Rousey, 01] par exemple le type de concepts $t_{c1.1.1.1}$ représente la notion définie par '' un système convertissant une énergie thématique en énergie mécanique'' (le moteur).
- Organiser les types de concepts dans une hiérarchie. Une notion se définit par les différences et les liens qu'elle entretient avec les autres notions. Par exemple, il faut différencier, '' un système propulsant un véhicule automobile '' (le type de concepts $t_{c1.1.1.2}$) du '' système propulsant un véhicule aérien '' (le type de concepts $t_{c1.1.1.3}$) même si ces deux types de systèmes convertissent l'énergie thermique en énergie mécanique ($t_{c1.1.1.2}$ spécialise $t_{c1.1.1.1}$ et $t_{c1.1.1.3}$ spécialise $t_{c1.1.1.1}$). On peut aussi employer le ''est'' et '' une partie de '' de relations.
Et enfin, une définition permet de normaliser sémantiquement une notion. Les définitions sont écrites en langage naturel pour faciliter la comparaison des notions par les utilisateurs potentiels du système. Et comme nous nous situons dans un contexte multilingue, il nous faudra une définition par langue d'une notion caractérisée par un type de concepts.

4.5.2.3. Hiérarchie des types de relations:

Une fois les notions du domaine normalisées et caractérisées par des types de concepts, il est aussi nécessaire de normaliser les liens sémantiques entre ces notions.

Cette normalisation aboutit à la définition de types de relations de l'ontologie (thésaurus sémantique). Chaque relation a un identifiant et une définition.
La normalisation des liens sémantiques est similaire à celle des notions. Cette normalisation se fait par l'intermédiaire d'une définition textuelle. La définition peut contenir, entre autres, la description sémantique du lien ainsi que son contexte d'utilisation. Un lien sémantique est identifié par un type de relations qui n'est pas un terme. Pour compléter leur définition, les types de relations sont organisés dans une hiérarchie de spécialisation.
La relation ''est un'' et la relation ''sorte de'' illustrent parfaitement la nécessité de la normalisation des relations, car selon les domaines ou les auteurs, ces relations ont une sémantique variable. Dans le langage objet, la relation ''est un'' correspond à la relation d'instanciation et associe une instance à sa classe, alors que la relation ''sorte de'' établit l'héritage entre classe. Cette dernière correspond à notre relation de spécialisation des types. Donc, de la même manière que les notions, les liens sémantiques ou relations ont besoin d'être normalisés, c'est à dire de fixer leur définition et leur interprétation.

4.5.2.4. Les relations entre types :

L'ensemble des types de concepts Tc et l'ensemble de types de relations T_R sont ordonnés par une relation de spécialisation, notée $\leq_c$ pour Tc et $\leq_R$ pour T_R. Dans le cas des types de concepts, dire que $t_{c1.1} \leq_c t_{c1}$ signifie que :

- Le type de concepts $tc_{1.1}$ est plus spécifique que le type de concept $tc_{1.}$
- $t_{c1.1}$ spécialise $t_{c1.}$
- Dans la hiérarchie des types de concepts, $t_{c1.1}$ est un descendant de $t_{c1.}$

La relation inverse de la relation de spécialisation $\leq_c$ notée $\geq_c$, est appelée relation de généralisation, si $t_{c1.1}$ spécialise tc_1 alors tc_1 généralise $t_{c1.1.}$

$$t_{c1.1} \leq_c t_{c1} \Leftrightarrow t_{c1} \geq_c t_{c1.1}$$

Par exemple, d'après l'ensemble des types de concepts de la figure 4.7, si $T_{c2.1}$ représente la notion de mammifère et $T_{c2.1.2}$ celle de vache, alors $T_{c2.1.2} \leq T_{c2.1}$ signifie que vache spécialise mammifère et $T_{c2.1} \geq T_{c2.1.2}$ signifie que mammifère généralise vache.

De la même manière, on définit pour les types de relations, la relation de généralisation, notée $\geq_R$.
Pour simplifier les explications ultérieures, nous introduisons la notion de comparabilité. Deux types sont dits comparables s'il existe une relation de spécialisation ou de généralisation entre eux.

4.5.2.5. Définition formelle du thésaurus sémantique:

Une ontologie (composée de *P* langues) est un quadruplet $M = (S, V, \lambda_C, \lambda_R)$ disposant de :
- $S = (T_C, T_R, \sigma)$, un support composé d'un ensemble de types de concept T_C, d'un ensemble de types de relation T_R, d'une application σ qui, à chaque type de relations, fait correspondre sa signature.
- *V,* un ensemble des vocabulaires, partitionné en ensembles de termes appartenant à la même langue (un vocabulaire). $V = V_{L1} \cup V_{L2} \cup \ldots \cup V_{Lj} \cup \ldots \cup V_{LP}$ où V_{Lj} est l'ensemble des termes appartenant à la langue *Lj*.

- $\lambda_C = \{\lambda_C^{VL1} \ldots\ldots \lambda_C^{VLj} \ldots\ldots \lambda_C^{VLP}\}$, un ensemble de P d'applications telles que λ_C^{VLj} : Tc $\rightarrow$ V_{Lj} est une application qui, à chaque type de concept $t_c \in T_C$, fait correspondre un terme dans la langue Lj, $\lambda_C^{VLj}(t_c) \in V_{Lj}$.
- $\lambda_R = \{\lambda_R^{VL1} \ldots\ldots \lambda_R^{VLj} \ldots\ldots \lambda_R^{VLP}\}$ Est un ensemble de P applications telles que λ_R^{VLJ} : T_R $\rightarrow$ V_{Lj} est une application qui à chaque type de relation, $t_r \in T_R$, fait correspondre un terme dans la langue Lj, $\lambda_R^{VLj}(t_r) \in V_{Lj}$.

Remarque : La fonction λ_C^{VLJ} ou λ_R^{VLJ} associe à chaque type de concept ou de relation un ensemble de termes appartenant au vocabulaire V_{Lj}. A cette fonction correspond une application de même nom qui à chaque type de concept ou de type de relation associe un seul et unique terme appartenant à un sous ensemble de V_{Lj}. Pour clarifier la formalisation, nous considérerons, par la suite, que les fonctions de λ_C ou λ_R sont des applications en restreignant leur ensemble d'arrivée V_{Lj}.

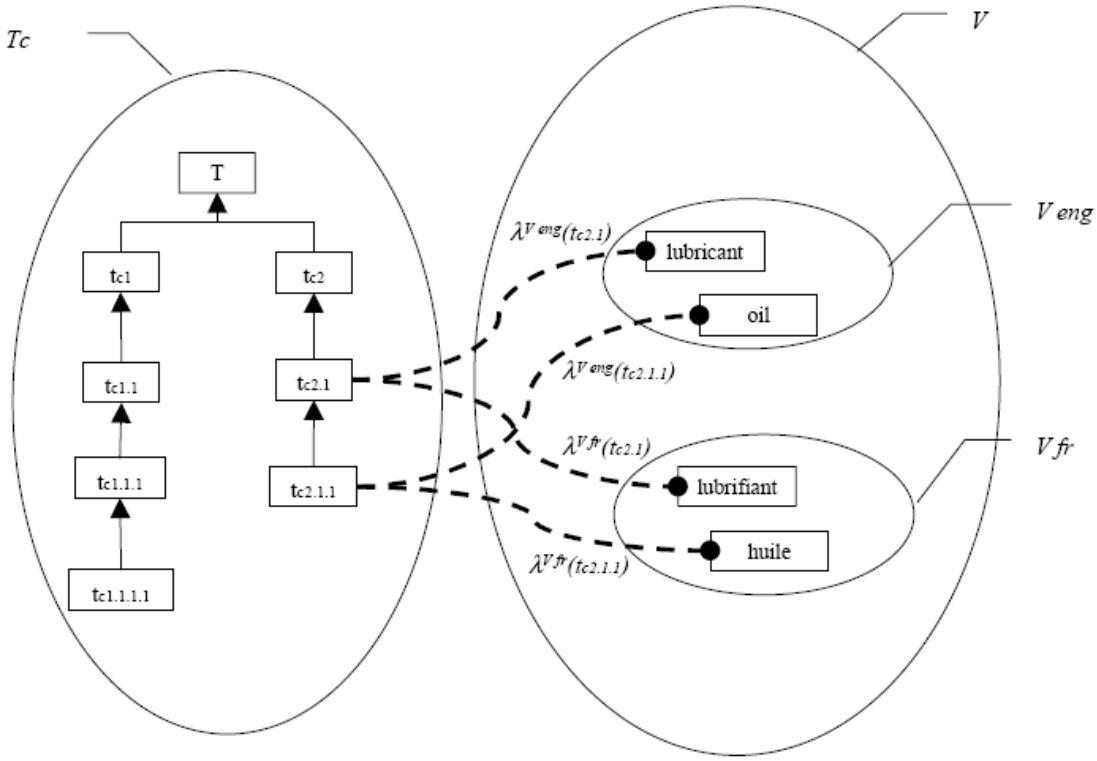


Figure 4.8 : Exemple de l'ontologie

La figure 4.8 présente un exemple d'application λ_C. Dans cet exemple, V se compose de deux vocabulaires, un vocabulaire anglais V_{eng} et un vocabulaire français V_{fr}. Pour tout type de concept, correspond un terme de chaque vocabulaire. Par exemple, pour le type $t_{c2.1} \in T_C$, $\lambda_C^{Veng}(t_{c2.1})$="lubricant" et $\lambda_C^{Vfr}(t_{c2.1})$="lubrifiant".

A partir de ce thésaurus sémantique définissant le vocabulaire et les connaissances du domaine, nous allons pouvoir définir les graphes sémantiques. La formalisation que nous proposons, a pour but de mettre l'accent sur les couples de concept reliés par une relation. Un exemple de couple de concepts est représenté dans la figure 4.9. C'est pourquoi dans le modèle des graphes sémantiques est définit la notion d'arc : un arc se compose d'un couple de concepts étiqueté par un type de relation.

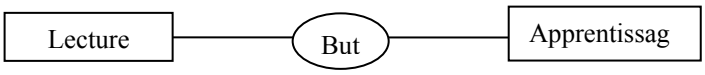


Figure 4.9 : un graphe conceptuel composé d'un couple de concepts relié par une relation représente 'La lecture pour l'apprentissage'

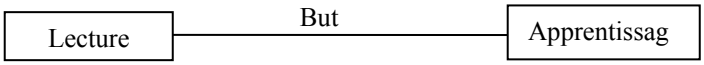


Figure 4.10 : un graphe sémantique composé d'un couple de concepts relié par un arc

La figure 4.9 et la figure 4.10 présente le même exemple d'arc, à l'aide du formalisme des graphes conceptuels pour la figure 4.9, et de formalisme des graphes sémantiques pour la figure 4.10.

En conclusion, dans un thesaurus sémantique les connaissances du domaine constituent le langage pivot, utilisé pour l'indexation, et les connaissances terminologiques contiennent les labels possibles, dans plusieurs langues, des connaissances du domaine. Ainsi, un graphe sémantique adapte sa présentation des connaissances en fonction d'une langue donnée. Donc nos index écrits à partir des connaissances du domaine, pourront être lisibles dans plusieurs langues à l'aide des connaissances terminologiques.

Comme nous avons présenté, dans la figure 4.6, notre système de recherche d'information multilingue se décompose en trois étapes :

4.6. Construction manuelle du Thésaurus sémantique :

La première étape dans l'élaboration de notre système est basée sur l'ensemble homme-machine. Les experts de domaine créent et mettent à jour le modèle du domaine (support) avec les vocabulaires associés à travers des interfaces interactives.
La conceptualisation du domaine alors ne pourra être modifiée que par les acteurs de l'indexation, et les vocabulaires évoluent en fonction de la terminologie des documents du corpus.

4.7. Indexation, extraction et génération des GS:

Le système d'indexation est basé sur des algorithmes d'extraction des connaissances à partir de documents textuels. L'approche que nous avons adoptée pour l'extraction est basée sur le calcul de segments (séquences) répétés. Le processus d'indexation distingue deux sortes de connaissances : la connaissance du domaine qui est indépendante de la langue et la connaissance terminologique ou les vocabulaires associés.
Un segment répété est une séquence de mots qui apparaît au moins deux fois dans un texte du corpus [Oueslati, 99] [Aliane et al, 06].

Les index sont mis en application en tant que graphes sémantiques qui sont construits à partir des textes selon les étapes suivantes :
1. Identification de la langue du document,
2. Le processus d'extraction est basé sur la méthode du calcul des segments répétés et produit en sortie une liste de candidats termes et une liste de candidats relations,

3. Cette liste est utilisée pour chercher des concepts et des relations dans l'ontologie et qui correspondent respectivement aux candidats termes et aux candidats relations,
4. Les concepts et les relations extraits à partir des étapes précédentes et qui ne sont pas trouvés dans l'ontologie seront ajoutés à cette ontologie ainsi qu'à l'index des documents.

Donc, les index des documents ou de la requête décrivent les concepts et les relations entre les concepts décrivant les relations et les termes contenus dans le texte et qui sont jugés intéressants pour le domaine, tout en se basant sur les concepts et les relations de l'ontologie.

Comme nous l'avons cité plus haut, l'approche standard de la modélisation d'un domaine à partir des textes consiste à

(i) identifier les concepts caractérisant le domaine, pour ensuite
(ii) Extraire les relations sémantiques qui les unissent.

Donc, nous proposons une approche pour la génération d'un graphe sémantique qui se décompose en deux étapes: (i) détection d'une instance d'un concept dans le texte constituant les arguments de relations, et (ii) identification des relations.

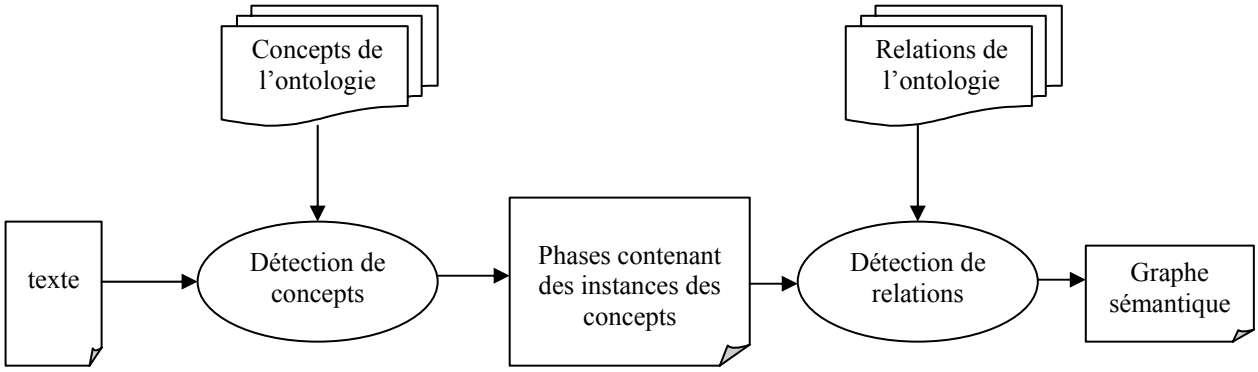


Figure 4.11 : les étapes de la génération d'un graphe sémantique basé à partir du texte

Dans ce qui suit, nous détaillons les différentes étapes de notre méthodologie présentée dans la figure 4.11.

Nous proposons de mettre en place une méthode d'extraction de termes appropriée à un domaine spécialisé. Nous avons fait le choix sur un domaine scientifique et technique (informatique). Cette méthode utilise, dans un premier temps, un extracteur terminologique et, dans un second temps, une méthode de filtrage.

4.7.1. Extraction des termes à partir des textes :

Dans cette section, nous développons les différents aspects de l'extraction de termes dont une partie a été vue dans le Chapitre 1, sous forme de quelques exemples.

4.7.1.1. Contexte :

L'extraction de termes est une étape fondamentale dans un processus de recherche d'information; une partie de la qualité des résultats en dépend. Pour cela, nous nous sommes basées sur les différentes approches d'extraction de termes, énumérées dans

le paragraphe suivant, qui s'appuient sur les présupposés suivants (ou sur une partie d'entre eux, selon la stratégie d'extraction adoptée) d'après [L'Homme, 01] :

1. les textes d'un corpus, dans le cas d'un domaine spécialisé, comportent un ensemble de termes qui caractérisent les connaissances spécialisées du domaine : les termes du domaine ;
2. un terme du domaine sera utilisé à plusieurs reprises dans un texte spécialisé : le terme possède une fréquence plus importante dans ce texte que les autres termes, et/ou le terme est plus fréquent dans ce texte spécialisé que dans un texte général ;
3. la plupart des termes sont des syntagmes nominaux ; i.e que la très grande majorité des termes sont de nature nominale.
4. la plupart de ces termes sont dits complexes, c'est-à-dire qu'ils sont composés de plusieurs mots. Ces mots sont par ailleurs utilisés isolément (ex. Intelligence artificielle, contrat de travail) ;
5. Les termes complexes se construisent au moyen d'un nombre fini de séquences de catégories grammaticales. En effet, la plupart des termes complexes français se composent d'un nom modifié par (autrement dit, il s'agit d'un sous-ensemble de syntagmes nominaux) :

- Un adjectif : ex. *intelligence artificielle, haute tension*;
- Un syntagme prépositionnel contenant un nom : ex. *robinet de commande, traitement de la demande*;
- Un syntagme prépositionnel contenant un verbe : ex. *machine à coudre*;
- Un autre nom : ex. *imprimante laser, page Web*;
- N'importe quelle combinaison des séquences ci-dessus : ex. *temps de conduction auriculaire.*

Pour faire identifier les termes complexes automatiquement, on fait appel à des techniques regroupées traditionnellement dans deux catégories : les stratégies *linguistiques* ou *statistiques*. Les premières associent des informations linguistiques à des chaînes de caractères ou font appel à des connaissances sur la langue traitée (au moins minimales) et recherchent, le plus souvent, des suites de catégories grammaticales. Les secondes effectuent des calculs sur les chaînes de caractères et s'appuient sur le fait que des termes significatifs sont employés forcément plus d'une fois dans un texte spécialisé. Pour résumer, ce principe veut que l'association récurrente de deux mots ne peut être que le fruit du hasard mais forcément significative. Concrètement, les occurrences des mots d'un texte sont examinées de la manière suivante : si un mot X apparaît plus fréquemment dans l'entourage d 'un mot Y qu'ailleurs dans le texte, alors X et Y forment une combinaison significative.

4.7.1.2. Analyse de surface pour l'extraction des syntagmes nominaux :

Partant de l'hypothèse qu'un groupe de mots qui occurrent ensemble dans un même contexte détermine des concepts appropriés, notre méthode d'extraction s'appuie sur l'extraction des syntagmes nominaux, puisque, plusieurs travaux ont montré le lien entre syntagmes nominaux et thèmes (*ce dont on parle* ou *ce dont il est question*), ils s'accordent sur le fait que seuls les groupes nominaux peuvent être des référents [Ama, 00][Haddad, 02][Aliane et al, 06].

De même, dans un processus de communication, le thème est considéré comme étant le point de départ de la communication et son support. C'est pourquoi dans le domaine de la RI les syntagmes nominaux ont eu plus d'attention.

Les résultats rapportés montrent que le regroupement statistique (basée sur la cooccurrence des mots) est meilleur que le regroupement syntaxique [Fagan, 87][Baziz, 05].

Pour une méthode statistique la langue dans laquelle les textes sont exprimés n'a que peu d'importance. Une conséquence importante est l'indépendance du notre système vis à vis de la langue utilisée dans les textes à traiter.

Nous avons ainsi choisi d'utiliser l'apprentissage pour acquérir les concepts correspondants aux textes traités. L'apprentissage automatique du langage par le comptage d'occurrences a déjà été étudié par Andreewsky Enguehard et al, 92] mais le but était de découvrir la grammaire de la langue.

Notre travail se destine donc à un corpus technique dans lequel les textes sont généralement écrits dans un langage dit « opératif », un langage précis comportant peu d'homographes ou de synonymes [Enguehard et al, 92] .

[Aliane et al, 06] a proposé d'utiliser pour l'enrichissement des graphes sémantiques, la méthode des segments répétés dans l'objectif de tester plus tard le système sur la bibliothèque du CERIST qui est constituée de documents en arabe, en français et en anglais. Nous avons essayé dans le travail que nous décrivons ci dessous d'améliorer cette approche pour répondre au mieux aux objectifs attendus.

La méthode des segments répétés s'appuie sur la détection de chaînes constituées de morceaux existant plus de deux fois dans le même texte [Haddad, 02]. Cette approche commence par stocker tous les mots du texte dans une table dont la valeur correspond soit à un terme, soit à une ponctuation, soit à un symbole de structure du texte (saut de paragraphe, chapitre, etc.) et une fréquence minimale d'apparition dans le texte est fixée afin d'éliminer les faibles occurrences. Pour chaque forme du texte, l'ensemble des suites dans le texte commençant par cette forme est répertorié. Le processus est réitéré pour chaque forme du texte.

Le système ANA [Enguehard et al, 92] se distingue par l'absence de pré-traitement des corpus. Les différentes étapes de la méthode sont décrites dans la figure suivante :

Données	Corpus de textes bruts
Familiarisation	Extraction automatique de quatre listes : mots fonctionnels, mots fortement liés, mots de schémas, bootstrap (termes du domaine)
Découverte	Extension de la liste des termes du domaine à partir du bootstrap en utilisant les patrons suivants : - Expression : terme constitué de deux termes co-occurrents. - Candidat : co-occurrence d'un terme, d'un mots de schéma et d'un mot (nouveau terme). - Expansion : co-occurrence d'un terme et d'un mot.
Résultat	Liste des termes

Figure 4.12 : Déroulement de la méthode ANA pour l'acquisition automatique de termes

ANA est mis pour *Apprentissage Naturel Automatique,* en raison d'une similitude établie par l'auteur entre l'apprentissage naturel, chez un enfant par exemple.
ANA est un système de construction automatique de thesaurus dans un domaine de spécialité, c'est-à-dire, selon l'auteur, *"un ensemble de termes et de syntagmes issus d'un domaine, ...normalisé et habituellement structuré à l'aide de lien typé"*. Un thesaurus se distingue d'une simple terminologie en ce qu'il est structuré. ANA procède donc à deux opérations, le repérage des concepts, et leur représentation dans un réseau sémantique.

4.7.1.3. Fonctionnement de notre extracteur de termes:

Le texte à traiter est nettoyé de tous les caractères de mise en page et de ponctuation.
Il est converti et exprimé dans un alphabet limité aux lettres latines minuscules et au caractère blanc pour ce qui est du français et anglais, de toutes les lettres pour l'arabe d'où le choix de lemmatisation.
La lemmatisation est l'opération qui consiste à réduire les formes fléchies des mots à leur racine grammaticale. Par exemple, les mots « voiture », « voitures » et « voituriers » auront pour lemme « voiture ».
Un mot peut être modélisé comme suit : préfixe + racine + suffixe [Moens, 00] [Christophe, 04]. L'objectif de la lemmatisation est d'améliorer les performances de notre extracteur de termes grâce à un appariement entre le vocabulaire des documents du corpus ou de la requête et le bootsrap. Cette opération permet de réduire le nombre de termes dans un index, ce qui est intéressant du point de vue du stockage des données.

Pour l'extraction des termes, on se base sur le postulat qui précise que les cooccurrences fréquentes sont significatives pour cela nous avons basé sur la méthode segments répétés.

- La méthode des segments répétés :

On stocke tous les mots du texte dans une table *tnum*(i) dont la valeur correspond à une occurrence. Pour chaque forme A du texte, n'étant pas ni un article, ni une conjonction, ni une préposition (c'est-à-dire les mots vides), ni un nombre, ni un verbe, on répertorie toutes ses occurrences dans un tableau *list*(j) (jième occurrence de A). La forme A sera appelée *tête de syntagme*. On cherche donc toutes les expansions (propagation) se rapportant à cette tête (trouver les expansions gauches ou droites).
On n'a seulement besoin de *tnum* et *list* pour identifier les suites. Chaque suite sera conservée seulement si sa fréquence est supérieur à 2.

Pour la recherche des expansions de la tête du syntagme, on se base sur le modèle du système ANA.
L'approche de ANA est basée sur le postulat « Les événements fréquents sont significatifs ».
Ce postulat peut être appliqué :
- pour rechercher des séquences de mots répétitives,
- pour identifier des configurations dénotant des concepts.
Ces configurations privilégiées sont implantées sous forme de deux modules symétriques que l'on tentera de faire correspondre avec le texte. Si l'on rencontre l'une de ces configurations :

mot inconnu	mot spécifiant un schéma	concept connu

concept connu	mot spécifiant un schéma	mot inconnu

Alors le mot inconnu est considéré comme susceptible de devenir un concept.
Les mots spécifiant les schémas sont acquis par apprentissage si le corpus est suffisamment important, ou donnés sous forme déclarative.

4.7.1.3.1. Les connaissances déclaratives :

- Liste de mots vides :

La liste de mots vides rassemble quelques prépositions, conjonctions, adverbes, etc qui sont considérés comme non significatifs. Ils ne pourront ni obtenir le statut de concept à titre individuel, ni figurer au début ou à la fin d'un concept.
Nos algorithmes de recherche des segments répétés ne tiennent pas en considération les mots vides. En effet, un mot vide (stopword en anglais) est par définition un mot non significatif dans un processus de recherche documentaire.
Pendant la phase d'indexation, on ne peut déterminer, a priori, les mots significatifs (ou mots-clef) de chaque document car chaque mot composant le document est un mot-clef potentiel.
Cependant, on peut éliminer les mots ayant certaines caractéristiques tels que par exemple les mots trop fréquents. Les articles et les déterminants en sont un exemple. Cette élimination, qui permet de réduire la taille de l'index, se justifie par le fait qu'ils sont présents dans la quasi-totalité des documents du corpus et ne peuvent ainsi être discriminant dans une requête. La prise en compte de ces mots dans une requête risque de retourner le corpus dans sa quasi-totalité. Les mots vides peuvent être également éliminés à l'aide d'une liste préalablement définie de mots (stop list) pour le français et l'anglais.

Contrairement aux langues latines, l'arabe est une langue agglutinante ; Les articles, les prépositions et les pronoms collent aux adjectifs, noms, verbes et particules auxquels ils se rapportent ; ce qui engendre une ambiguïté morphologique au cours de l'analyse des mots. De même tous ces mots vides de l'arabe peuvent être concaténés entre eux.
Par exemple de 'تلك ' on peut dériver 'فبتلك' = تلـــك + ب+ف et. تلـــك =وتلك +و
Donc, pour traiter le phénomène de concaténation on considère la racine de ces mots. D'où une liste des formes canoniques est utilisée et des algorithmes de dérivation sont implémentés.

- Les schémas :

Puisque notre but c'est bien l'extraction des syntagmes nominaux, la structure de ces syntagmes est composée souvent par des mots fonctionnels (mots de schéma).
Dans le cas du français, il y a généralement moins de 10 éléments dans cette liste. Exemple {"de", "de la", "des", "du" "en"}.

- Les verbes :

Pour l'extraction des mots qui ne sont pas des verbes, on peut ajouter la liste des verbes existant dans le corpus (un lexique de verbes).

4.7.1.3.2. Induction et extraction de nouveaux termes :

Le postulat précise que les cooccurrences fréquentes sont significatives [Enguehard et al, 92]. Lors de l'induction de nouveaux termes, notre système passe en revue les objets susceptibles de mémoriser plusieurs occurrences identiques. Si une forme figée (toujours en suivant les critères de l'égalité souple) est assez fréquente, le système crée un nouveau terme.
Techniquement, le texte est vu à travers d'une fenêtre de 2 à 10 mots. Les mots vides et ceux de moins de deux lettres ne sont pas pris en compte dans le calcul de l'empan de cette fenêtre.
La fenêtre est déplacée tout le long du texte, son contenu est recueilli suivant trois voies différentes en fonction de sa nature.

- Cas 1 :
Lorsque le système voit deux concepts, il note l'occurrence, c'est à dire l'extrait de texte que laisse voir la fenêtre, dans un objet du type "expression" particulier ces deux concepts,
Exemple : Soit le texte : *"je voudrais un VERRE d'EAU ou de ..."*
‘’ أريد كأسا من الماء أو من ‘’
L'occurrence *"VERRE d EAU"* est écrite dans l'objet expression correspondant.

- Cas 2 :
S'il ne voit qu'un concept (ici "VERRE"), le contexte local est analysé pour repérer un schéma, et donc un mot potentiellement intéressant ("lait"). Un objet de type candidat portant son nom recueille l'occurrence.
Exemple : Soit le texte *"j'ai réservé mon VERRE de lait devant.,."*

L'occurrence *"VERRE de lait"* est écrite dans le candidat "lait".

- Cas 3 :
Si l'examen du contexte local ne fait apparaître aucun schéma connu, l'occurrence est également conservée dans un champ spécifique. Elle sera traité différemment.
Exemple : Soit le texte *"Voici de I'EAU minérale"* L'occurrence *"Voici de I'EAU minérale"* est écrite dans le candidat "EAU". En arabe ‘’ الطبيعـــــي المـاء من هذا ‘’
Donc, le texte est balayé de la gauche vers la droite. Toutes les cooccurrences d'événements de ces trois types 'expression' ou 'candidat' ou 'expansion' sont stockées dans des objets de la base de données.
Le processus de traitement d'un texte s'arrête lorsque aucun nouveau terme n'apparaît pendant un cycle.

4.7.1.3.3. Analyse des occurrences :

Cette phase de lecture est suivie de I'examen des informations recueillies. Seuls les objets ayant recueilli plus de 2 occurrences sont examines.

- Les expressions :

Si la même configuration, aux variations morpho-syntaxiques prés, se présente 2 fois au moins, elle devient un concept sous sa forme la plus fréquente,
Exemple : Voici les occurrences de l'expression rassemblant "VERRE" et "EAU" :
"Je voudrais un VERRE d'EAU ou de..."
"Bois un VERRE d'EAU pour faire..."
"Aspirine dans ton VERRE d'EAU..,"
L'analyse va qualifier le nouveau concept "VERRE D'EAU"

- Les candidats et les schémas

Les candidats dont la fréquence est supérieure au seuil m deviennent eux-mêmes des concepts sous la forme morpho-syntaxique la plus fréquente.
Exemple : Voici les occurrences du candidat "lait" :
"J'ai renversé mon POT de lait devant.,."
"Distribuer un VERRE de lait à chacun.,."
"Boire un VERRE de lait c'est..."
"Je préfère un VERRE de lait nature., "
"J'ai vidé la BOUTEILLE de lait qui était.''
L'analyse va qualifier le nouveau concept "LAIT"

- Les candidats sans schéma

Les concepts existants présentant n fois le même contexte local engendrent un nouveau concept intégrant ce contexte.
Exemple : Voici les occurrences sans schéma du candidat "VERRE" :
"Bois un grand VERRE cela ira mieux,.."
'7'ai acheté un VERRE de café ..."
"Voici ce grand VERRE dont je t'ai parlé..."
L'analyse va qualifier le nouveau concept "GRAND VERRE"

La liste des candidats termes extraits sera employée pour faire la recherche des concepts de l'ontologie.
Cette phase se limite à la reconnaissance des concepts connus, c'est à dire les termes de l'ontologie sont identifiés (grâce aux procédures d'égalité souple). A ce stade, le texte est perçu comme une liste de termes connu et de termes inconnu. Les nouveaux termes sont ajoutés à l'ontologie grâce à une validation humaine.

4.7.1.4. Expansion de la liste des candidats-termes :

Comme nous pouvons le remarquer, la phase d'extraction de termes telle qu'elle est décrite dans les paragraphes précédents se base essentiellement sur les termes répertoriés dans l'ontologie. C'est-à-dire que les documents sont indexés suivant les concepts de l'ontologie. Un des problèmes rencontrés vient du fait qu'un certain nombre de concepts (relatifs au domaine) extraits de ces documents, ne sont pas présents dans l'ontologie peut être adapté au domaine particulier de l'utilisateur.
La liste des termes inconnus extraits des textes sera ajoutée à l'ontologie suite à par une validation de l'expert à travers une interface de mise à jour de l'ontologie.

4.7.2. Démarche de l'extraction des relations sémantiques :

Nous nous basons sur l'idée: '' le sens d'un mot est lié à ses contextes d'utilisation'' pour proposer notre approche d'extraction des relations sémantiques.
Comme nous avons cité, dans la partie d'extraction des connaissances du chapitre 1, l'analyse statistique consiste à lier les termes qui partagent des contextes identiques ou bien similaires. Le contexte statistique, souvent appelé fenêtre de mots est composé de l'ensemble des termes entourant le terme en question dans le texte en question.
Dans le cadre de notre travail, nous nous sommes basé sur l'analyse statistique pour lier les termes (syntagmes nominaux) puisque l'idée de cette analyse est que peuvent être rapprochés des termes qui, non pas cooccurrent souvent, mais qui apparaissent souvent avec des contextes similaires. Ainsi deux termes apparaissant souvent dans les fenêtres des mots, ou contextes, proches seront reliés, ces deux termes n'apparaissant pas forcément dans les mêmes documents. Une telle méthode est décrite notamment dans [GWR, 99][Lame, 02]. Le choix de la longueur de la fenêtre de mot est fondamental dans cette méthode. Plus ce contexte, une phrase, un paragraphe ou un document, est spécifique, plus les relations établies entre les termes seront précises.
L'analyse statistique est basée sur hypothèse [Rijsbergen, 79] : *« L'emploi de 2 termes en cooccurrence est l'expression d'une relation sémantique entre eux ».*

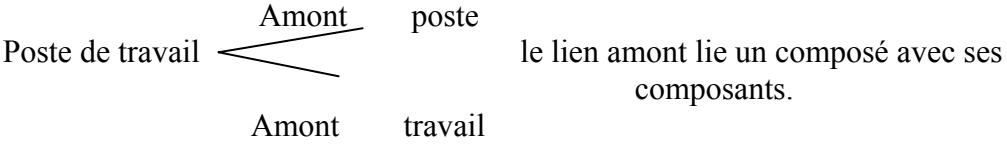
Comme nous avons vu, dans la partie d'extraction des connaissances du chapitre 1, nous avons défini deux types de relations sémantiques :

4.7.2.1. Les relations syntagmatiques :

Sont des relations de co-occurrence direct entre deux termes. Pour détecter ce genre de relations, nous utilisons une fenêtre de 2 à 10 mots, puisqu'un terme extrait peut être un terme simple ou un terme composé.

D'après la phase d'extraction des termes, nous avons remarqué qu'il existe deux types de termes : termes simples composés d'un seul mot et termes composés qui se composent de deux ou plusieurs mots c'est-à-dire un ou plusieurs termes simples.
A partir de ces types de termes, on peut déduire en premier lieu un type de relation d'inclusion qui lie les mots (termes simples) d'un terme composé les uns des autres et avec le terme composé en question. Notre idée est basée sur l'approche du système ANA.

Exemple :
''Poste de travail'' est un terme composé de termes simples ''poste'' et ''travail'' et qui sont relié par le mot de schéma ''de''. Donc, nous avons un classement ou une collocation.



A ce niveau, on se base sur la relation syntagmatique, c'est une relation de co-occurrence directe entre deux termes.

4.7.2.2. Relations paradigmatiques :

Concernant les relations paradigmatiques (co-occurrence indirect) on se base sur l'idée '' deux termes sont similaires s'ils apparaissent dans des contextes similaires''. Pour cela on se base sur le modèle vectoriel qui suppose que les mots sont indépendants les uns des autres, c'est-à-dire qu'il n'y a pas d'information sur la position des termes.

Et comme notre corpus de textes comprenant un vocabulaire spécialisé technique, nous nous sommes basé sur les travaux de [GWR, 99] [Lame, 02]. Puisque la méthode de [Lame, 02] proposée a donnée des résultats acceptables sur tout lorsqu'elle est appliquée sur des corpus de textes spécialisés.
Cette méthode permet de détecter des relations de divers degrés entre des termes (synonymie, termes spécifiques, termes génériques, etc.…) mais qu'en aucun cas la nature de ces relations n'est identifiée.
Les étapes de cette méthode sont détaillées comme suit :

- Identification des termes cibles *w*. tout d'abord, doit être définie une liste de termes, termes cibles, dont les contextes d'apparition sont comparés. Nous utilisons pour ce faire la liste des termes extrais dans la première phase. Nous appliquons donc cette méthode de détection de relation entre termes sur les termes simples et composés.
- Définition du contexte : pour chacun des termes cibles identifiés, son vecteur de contexte doit être établi. Nous définissons les contextes des termes comme les documents. La liste des termes détectés comme tels par notre outil d'extraction de termes apparaissant dans le même article (document) que le terme cible en question constitue le vecteur des termes contextuels *c* (context words).
- Pondération des termes contextuels : chaque terme contextuel est pondéré sur la base de l'indice d'information mutuelle MI définit dans le chapitre 1 entre le terme contextuel et le terme cible en question. L'idée d'information mutuelle entre un terme cible et un terme contextuel est élevé quand leur fréquence conjointe f_{cw} est élevée relativement à leurs fréquences individuelles dans le corpus, f_c et f_w. La formule de l'information mutuelle ajoute 1 au ratio des fréquences de façon à ce que, à des occurrences de zéro correspondent des indices d'information mutuelle de zéro. La formule de l'indice d'information mutuelle [GWR, 99][Lame, 02] s'établit comme suit :

$$MI_{cw} = \log_2 \left[\frac{f_{cw}}{f_c \text{ et } f_w} + 1 \right]$$

- Comparaison des contextes : les vecteurs des termes contextuels pondérés sont alors comparés deux à deux pour chacun des termes cibles. La comparaison entre les vecteurs se fait par le biais de la mesure de similarité cosinus [GWR, 99][Lame, 02] qui peut varier entre –1 et 1. Plus la mesure de similarité cosinus est proche de 1 en valeur absolue, plus les vecteurs sont proches. La formule de la mesure de similarité cosinus entre les vecteurs de termes contextuels des termes a et b s'établit comme suit (p_a correspond aux poids des termes contextuels du terme cible *a*, p_b aux poids des termes contextuels du terme cible *b*, *ab* est le nombre de termes contextuels communs à *a* et à *b*) :

$$Sim_{ab} = \frac{\sum_{ab} p_a p_b}{\sqrt{\sum_a p_a^2 \sum_b p_b^2}}$$

- Détermination d'un seuil de similarité entre termes cibles. Dés lors que les indices de similarité entre chaque paire de termes cibles définis sont calculés, il faut déterminer le seuil à partir duquel les relations sont jugées valides. De la même façon que pour la méthode précédente, cette détermination ne peut se faire qu'en adoptant une méthode empirique. Cette méthode consiste à interpréter et à valider les relations entre termes détectées par divers niveaux des indices de similarité. Le seuil de similarité a été dans notre contexte fixé à 0.8.

Cette dernière phase d'extraction des relations sémantiques est soumise à une validation centrée utilisateur (expert), totalement assistée par un environnement informatique. C'est à dire que l'utilisateur soit valide une suggestion de relation entre deux termes (suggestion proposée par la phase de détection), soit, rajoute une relation présente dans le texte mais qui n'est pas suggérée par les phases de détection.

Une suggestion est un triplet constitué d'une relation et deux termes (terme1, relation, terme2) extraits à partir d'une phrase. Cette suggestion n'est valide que si la relation désigne bien la relation sémantique, et les deux termes sont bien liés par cette relation selon le sens du texte. Notons que les phrases présentées aux experts présentent une ou plusieurs suggestions.
La validation des suggestions par l'utilisateur (via des interfaces dédiées) assurant ainsi la qualité des graphes stockés dans la base de connaissances.

4.7.3. Génération de graphe sémantique :

Une fois les relations détectées et les termes extraits, la dernière phase consiste à identifier les désignations de chaque relation ainsi que les termes constituant les arguments de chaque relation afin de générer un graphe sémantique pour tout le document.

Comme nous avons cité, la validation des relations par l'utilisateur (via des interfaces dédiées) assurant ainsi la qualité des graphes stockés dans la base de connaissances.

Disposant de toutes ces informations, on applique alors l'algorithme de génération d'un graphe sémantique et qui se décompose en étapes suivantes:

0- cette phase consiste à effectuer une série d'analyses statistiques et filtrages sur le texte afin de le préparer pour les phases d'extraction.
1- La première phase consiste à analyser et traiter le texte pour l'extraction des syntagmes nominaux et constituer une liste de candidats termes. On utilise un extracteur suivant la langue traitée puisque dans notre cas nous avons plusieurs langues.
2- La deuxième consiste à faire la correspondance entre la liste des candidats termes et les termes de l'ontologie afin de filtrer la première et de n'en garder que les instances des concepts de l'ontologie.

3- Le but de cette phase est de repérer dans le texte, les relations sémantiques déjà modélisées dans l'ontologie ainsi que l'identification des noms de chaque relation. Passer à l'étape (4).
4- Vérifier que la relation retenue a été détectée comme étant une instance de relation de l'ontologie. Si c'est le cas, passer à l'étape (5), sinon récupérer la relation suivante et revenir à l'étape (3)
5- Vérifier que les arguments retenus vérifient bien la signature de la relation déjà définie dans l'ontologie. Si c'est le cas, passer à l'étape (6), sinon récupérer la relation suivante et revenir à l'étape (3)
6- Générer un graphe sémantique décrivant la relation. Récupérer la relation suivante et revenir à l'étape (1); s'il n'y a plus de relations générer le graphe sémantique global du document et le stocker dans la base de données contenant les index grâce aux algorithmes et FIN.

4.8. La recherche, étendre la requête via une ontologie :

L'utilisateur exprime sa question dans l'une des trois langues naturelles. Cette requête est analysée et transformée en graphe sémantique. L'ontologie de domaine est utilisée pour étendre la requête initiale dans l'objectif de trouver plus de documents pertinents.
Le processus de modélisation de la requête est le même que le processus de modélisation utilisé pendant l'indexation des documents.
Le processus de recherche consiste alors en une comparaison des graphes sémantiques pour trouver les documents qui répondent à la requête étendue d'utilisateur.

- Reformulation de la requête :

La qualité des réponses de système de recherche d'information dépend bien sûr de la qualité du mécanisme d'appariement requête/documents, mais aussi, de façon non négligeable, des requêtes formulées par l'utilisateur.
La lacune majeure des requêtes est leur manque de contexte, souvent créateur d'ambiguïté. En effet un besoin d'information est souvent exprimé par quelques mots. Donc, une autre manière d'améliorer les résultats est d'étendre la requête avec de nouveaux termes qui vont permettre de clarifier le concept caché derrière les termes.
Les ontologies peuvent être impliquées pour la modification de requêtes dans un SRI. L'objectif de notre approche est d'améliorer le rendement des requêtes en élargissant leur champ de recherche. La démarche suivie pour notre part, est d'abord de détecter les termes de la requête qui renvoient à des concepts de l'ontologie, puis, de les étendre par des termes représentant d'autres concepts proches de ceux de la requête. Ces termes sont identifiés grâce aux liens sémantiques entre concepts qu'offre l'ontologie.

Nous présentons dans ce qui suit une stratégie de modification automatique de requêtes par expansion guidée par ontologie, permettant une amélioration de la précision. Cette stratégie est basée sur une expansion sélective et prudente des concepts (mono ou multitermes) présents dans les requêtes.

L'ontologie peut intervenir dans le processus de modification ou de reformulation de requêtes en RI. L'idée, comme représentée dans la Figure 4.13, est de faire passer la

requête utilisateur par le réseau conceptuel d'une ontologie pour l'enrichir avec de nouveaux termes issus du vocabulaire de cette ontologie. L'intérêt est double :

- Permettre à l'utilisateur de faire usage d'une terminologie autre que celle présente dans les documents, donc de formuler ses requêtes dans son propre langage. C'est l'ontologie qui assure la correspondance entre les termes de l'utilisateur et ceux des documents.
- Permettre au SRI d'aider l'utilisateur à reformuler ses requêtes sur la base d'une proximité sémantique (processus automatique).

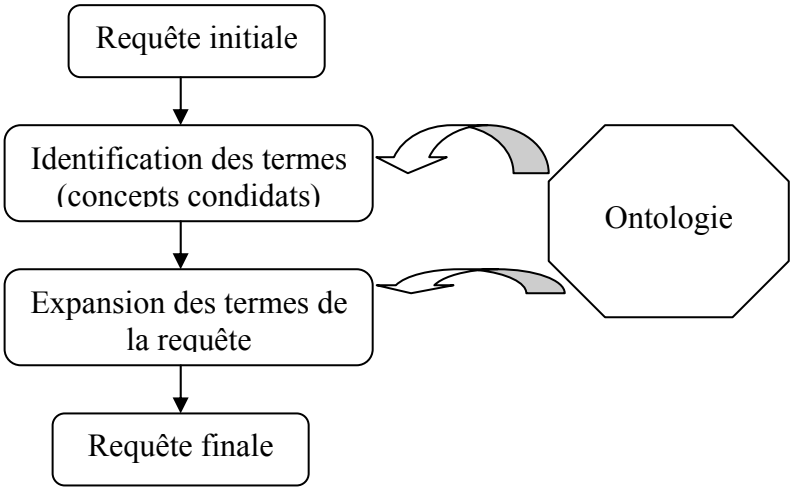


Figure 4.13 : cheminement d'une requête via l'ontologie

Globalement, l'approche de modification de requête proposée ici permet, étant donné une requête initiale, de la modifier en ajoutant des termes extraits de l'ontologie. Les termes ainsi rajoutés représentent des concepts liés à ceux de la requête initiale via différentes relations sémantiques (spécialisation/généralisation).

Pour réaliser ces modifications à la requête, il faut d'abord identifier les termes de la requête susceptibles de représenter des concepts dans l'ontologie. L'identification de terme adoptée ici suit tout comme pour les approches précédentes, la méthode décrite à la section 4.6. Les termes de la requête, une fois identifiés, l'affectation des concepts de l'ontologie associés est contextuelle, c'est-à-dire, pour affecter un concept à un terme de la requête, on prend en compte tous les autres termes de la requête.

Une fois le concept approprié est sélectionné, on peut rajouter à la requête les autres concepts reliés au concept sélectionné via les différentes relations sémantiques (spécialisation / généralisation).

[Aliane et al, 06] propose une approche heuristique pour étendre la requête en ajoutant le père le fils d'un concept de la requête avec des pondérations de similarité VRG et VRS respectivement. VRG et VRS sont deux constants arbitraires qui calculent la similarité de graphes.

Par exemple si une requête contient un seul concept $C_{1.1}$, l'expansion de cette dernière consiste à ajouter à la requête le père et le fils de avec des pondérations de similarité.

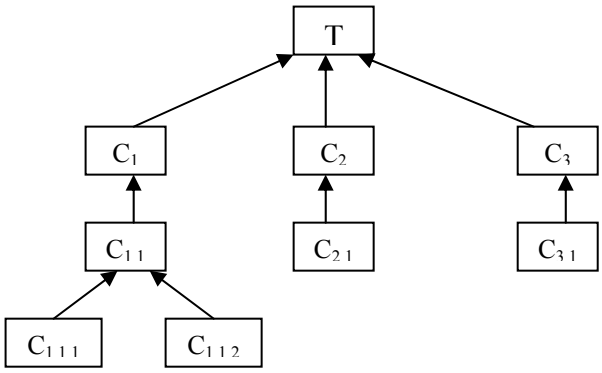


Figure 4.14 : Hiérarchie de type de concepts

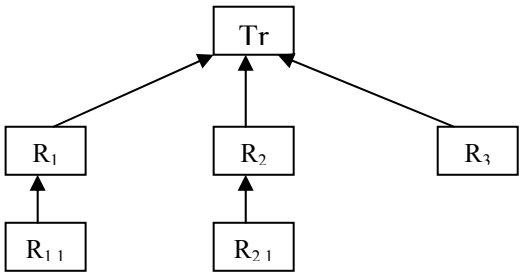


Figure 4.15 : Hiérarchie de type de relations

Soit le graphe sémantique de la requête G:

L'expansion du graphe G de l'exemple donne la requête :
Q'= {(C_1, 1), ($C_{1.1}$, 0.7), ($C_{2.1}$, 1), (C_2, 09), (C_3,1), ($C_{3.1}$,07), (R_1, C_1, $C_{2.1}$, 1), ($R_{1.1}$, C_1, $C_{2.1}$, 0.7), ($R_{2.1}$, C_3, $C_{2.1}$, 1), (R_2, C_3, $C_{2.1}$, 0.9) }.
Cet ensemble est utilisé pour rechercher les documents qui correspondent à chaque élément de cet ensemble.

Nous proposons une autre approche inspirée de [Baziz, 05] pour réaliser une modification optimale de la requête. [Baziz, 05] a prouvé que les résultats d'utilisation de différents types de relations sémantiques dénotant la spécialisation-généralisation ont montré que la méthode de modification de la requête par expansion "prudente" donne les meilleurs résultats en terme de précision.

La définition de la méthode d'expansion "prudente" est décrit comme suit :

Soit une relation appartenant à l'ensemble R={ R_s, R_g } des relations de l'ontologie et un terme T_k de Q_c. telque R_s est une relation de spécialisation et R_g est une relation de généralisation.

Premièrement, chaque terme de la requête est étendu en utilisant les différentes relations sémantiques. Le résultat de cette expansion est un ensemble de concepts candidats (concepts possibles) reliés au terme T_k:

$$R_i(T_k) = \{ C^1{}_k , C^2{}_k , ..., C^t{}_k \}$$

La manière d'étendre serait de n'utiliser que les "meilleurs concepts" de R^n pour étendre la requête. Dans ce cas, les meilleurs concepts seraient ceux ayant un taux de recouvrement (au sens défini par Lesk [Lesk, 88][Baziz, 05]) assez important avec la requête initiale. Cette désambiguïsation contextuelle permet de sélectionner les concepts en prenant en compte tous les termes de la requête.

La méthode d'expansion consiste à sélectionner pour chaque relation, donc pour chaque $R_i(T_k)$, le meilleur concept :

$$Best(R_i(C_k))=ArgMax_{k,j}||\{C^1{}_k, C^2{}_k, \ldots., C^j{}_k, \ldots., C^t{}_k\} \cap Q||$$

Dans ce cas, étant donné une relation R_i, le nombre de concepts à rajouter est égal au nombre de termes dans la requête. Cette façon d'étendre traduit l'hypothèse qu'une même requête peut faire allusion à plusieurs concepts différents.

La requête finale serait alors comme suit :

$Q' = \{ Q \cup Best(R_i(T_k)) \}\ i \in \{1, 2\}, k \in \{1, \ldots.,m\}$

Elle est représentée par les termes de la requête initiale, auxquels sont rajoutés les meilleurs concepts issus des différentes relations R_i.

Lors de la sélection, pour une relation donnée, si plusieurs concepts ont le même score (nombre de mots en commun avec la requête initiale), ceux ayant le plus grand nombre de mots différents prévalent. Viennent ensuite ceux ayant la plus grande taille (en nombre de mots).

En conclusion, l'idée est de tester les deux approches lors de l'implémentation, du système.

4.9. Architecture du SRIM basée sur un thésaurus sémantique :

Notre approche est résumée dans la figure 4.16.

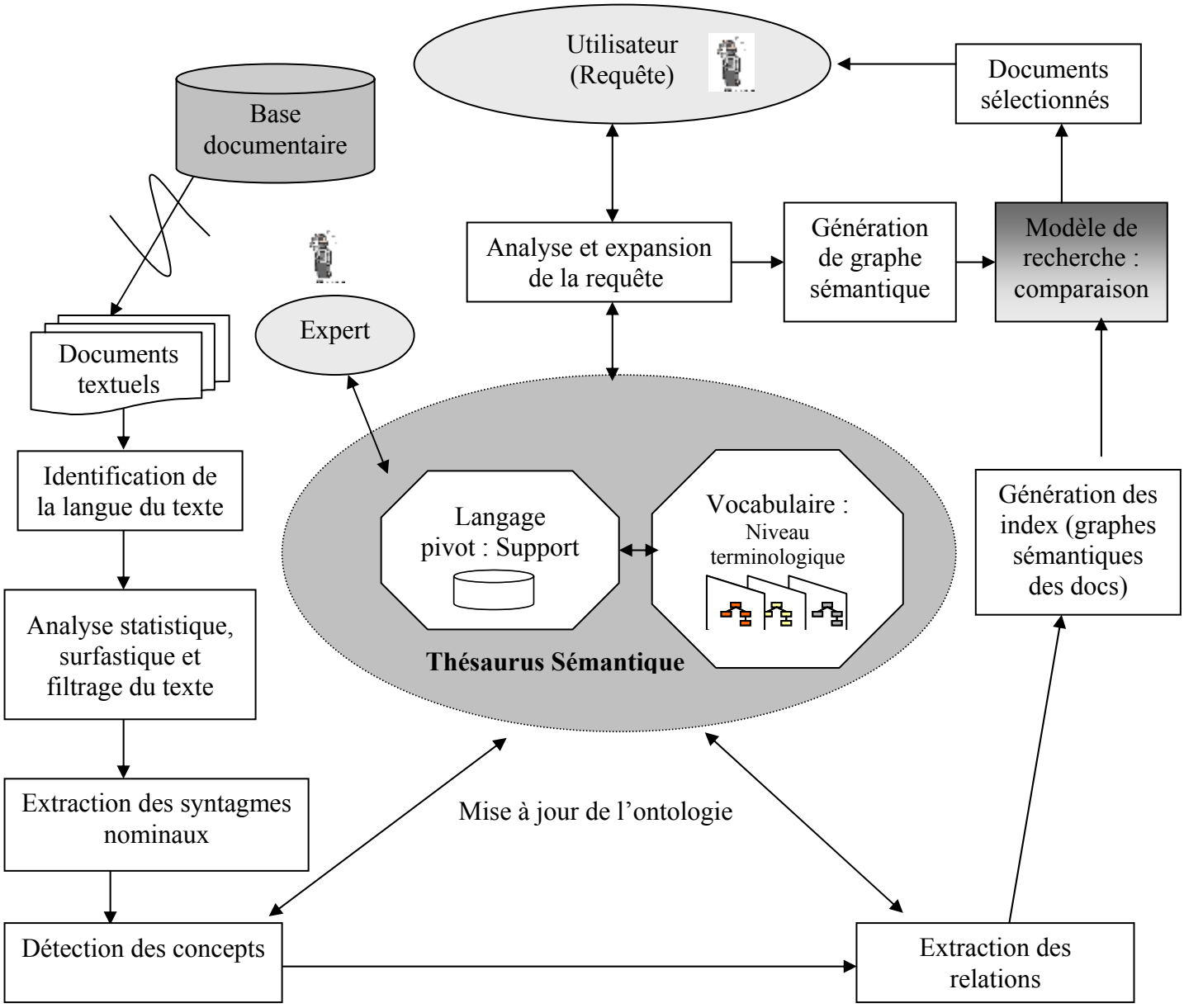


Figure 4.16 : Architecture d'un SRIM basée sur Thésaurus sémantique

4.10. Conclusion :

Nous avons présenté dans ce chapitre une architecture basée sur l'ontologie de domaine pour la recherche d'information multilingue. Cette ontologie, qui représente les connaissances du domaine, est fondée sur le formalisme des graphes sémantiques afin d'améliorer la sémantique des documents.
Dans un thesaurus sémantique les connaissances du domaine constituent le langage pivot, utilisé pour l'indexation, et les connaissances terminologiques contiennent les labels possibles, dans plusieurs langues, des connaissances du domaine. Ainsi, un graphe sémantique adapte sa présentation des connaissances en fonction d'une langue donnée. Donc nos index écrits à partir des connaissances du domaine, pourront être lisibles dans plusieurs langues à l'aide des connaissances terminologiques.

Nous avons proposé une méthodologie pour la génération (semi)-automatique des graphes sémantiques. Cette méthodologie décrit les relations entre les termes contenus dans le texte et qui sont jugés intéressants pour le domaine, tout en se basant sur les concepts et les relations de l'ontologie définit. En effet, pour la création des graphes, nous avons proposé d'utiliser les connaissances contenues dans les textes en se basant sur une méthodologie d'extraction de connaissances textuelles. Cette méthodologie est basée sur les techniques statistiques et de filtrage. Donc, l'extraction de connaissances permet d'indexer des nouveaux documents ainsi que la mise à jour de l'ontologie.

Conclusion Générale

Notre travail se situe dans le contexte de la recherche d'information multilingue : il faut retrouver tous les documents relatifs à une requête donnée quelque soit leur langue.

Notre approche est centrée autour d'une ontologie du domaine qui est fondée sur le formalisme de graphes sémantiques.

La première étape de notre système est une construction manuelle de l'ontologie (Thésaurus sémantique). La caractéristique du système consiste en une conception basée sur l'interaction homme-machine. En effet, le rôle de l'expert humain est primordial pour atteindre les objectifs de performance du système et cela en tant que gestionnaire de ressources de connaissances qui peuvent varier en « intelligence » ou en « puissance ». Dans cette étape, des interfaces hautement visuelles sont offertes pour l'expert humain en vue de la construction de son thésaurus sémantique. Le système est indépendant du domaine dans le sens où différents experts peuvent créer différentes bases de connaissances. Donc, la conceptualisation du domaine ne pourra être modifiée que par les acteurs de l'indexation, et les vocabulaires évoluent en fonction de la terminologie des documents du corpus.

Nous avons fondé cette ontologie sur le formalisme de graphes sémantiques introduits par [Roussey, 01] pour améliorer la sémantique des documents dans des SRIM.

Les experts créent, gèrent et mettent à jour cette ontologie.

Les index des documents et des requêtes sont représentés par les graphes sémantiques.

Les étiquettes sur les nœuds des graphes sémantiques sont des numéraux, de sorte qu'ils ne sont associés à aucune langue naturelle.

Nous avons proposé une approche d'extraction des connaissances basée sur une analyse linguistique de surface à savoir la méthode de segments répétés qui est utilisée pour indexer les nouveaux documents et mettre à jour l'ontologie.
Cette approche permet de traiter des documents textuels et d'extraire des instances de l'ontologie de domaine, ce qui contribue à la création de graphe sémantique comme index facilitant ainsi la recherche d'information.

Une approche d'expansion de la requête a été proposée en vue d'assurer une meilleure pertinence des réponses.

Ce travail doit faire l'objet d'une implémentation et d'une évaluation sur la bibliothèque du CERIST.

Comme suite à ce travail, nous pouvons envisager les perspectives suivantes :

- Une implémentation de l'approche proposée.

- L'approche d'extraction des relations sémantiques : dans notre travail, nous avons optée d'utiliser au niveau de la pondération des termes contextuels l'information mutuelle. C'est à dire que chaque terme contextuel est pondéré sur la base de l'indice d'information mutuelle MI, d'autres approches peuvent être envisagées.
- Comparaison des contextes : les vecteurs des termes contextuels pondérés sont comparés deux à deux pour chacun des termes cibles. Dans notre travail, La comparaison entre ces vecteurs se fait par le biais de la mesure de similarité cosinus, d'autres mesures peuvent être envisagées.
- La recherche : La prise en compte de la recherche dans notre système en intégrant le processus de recherche pour compléter le système.

Annexe :

1. Les connaissances déclaratives : Liste de mots vides :

Pour le français par exemple, cette liste est typiquement, des articles, des pronoms, quelques adverbes. Notre système les sélectionne automatiquement grâce à une procédure entièrement statistique. Cette liste comprend une centaine d'éléments. Exemple {"a", "alors", "après", "au", "auraient", "aussi", "autre", "avait", "avant", "avec", "avoir", "beaucoup", "c", "car", "ce", "cela", "celles", "certain", "ces", "cette", "ceux", "chacun", "chaque", "comme", "comment", "d", "dans", "de", "déjà", "des", "dirais", "dire", "dit", "donc", "du", "elle", "en", "encore", "est", "et", "était", "été", "eux", "il", "ils", "j", "je", "l", "la", "le", "les", "lors", "lui", "mais", "me", "même", "mêmes", "n", "ne", "non", "nous", "on", "ont", "par", "parce", "pas", "peu", "plus", "pour", "pouvait", "puis", "qu", "quand", "que", "quel", "qui", "s", "sait", "se", "son", "sont", "sur", "telle", "toujours", "tout", "toute", "toutes", "très", "trop", "un", "une", "vous", "vraiment", "y"}.

Pour la langue arabe, c'est comme le français, la liste des mots vides contient des patricules (الحروف), des pronoms (الضـــمائر) , des démonstratifs (أسـماء الاشـارة), des conditionnel(اسـماء الشـرط), etc.

La patricule est un mot invariable qui accompagne un nom ou un verbe, et ne peut véhiculer aucun sens quand elles sont isolées. Parmi les patricules, les uns s'emploient avec le nom, les autres avec les verbes et d'autres avec le nom et le verbe.

Dans l'étude des mots de la langue arabe, on distingue des patricules de :

1. Affirmatives (حروف التوكيد) : نعـم ,اجل
2. de Négation (حروف النفيـــــا) : كـلا , لا ,لـم , etc.
3. d'adjonction (حروف العطـف) : ثـم ,و ,ف, etc.
4. interrogatives (حروف الاســـتفهام) : هل ,أ ,مـتى, etc.
5. exclamatives (حروف التعجـــب) : ألا ,رب ,ربمـا , etc.
6. prépositions dites (حروف الجر) : الـى ، فـي، ب, etc.
7. conditionnelles (حـروف الشـرط): ان، لـو, etc.
8. d'inaccompli mansoub (حـروف النصـب) : أن ، لـن, etc.
9. d'inaccompli madjzoum (الجـزم حروف) : لـم ,لمـا, etc.
10. de détermination (حـــروف التعـــريف) : c'est l'article ال

cette liste contient aussi des :

11. démonstratifs (أسماء الاشارة) : désignent une ou plusieurs personnes (animaux ou choses), ils peuvent être soit pronoms soit adjectifs. Comme par exemple هنا, هذا ,تلـــك ’ذلـك ,هؤلاء ,أولائـــك etc.
12. Conjoints (الاســماء الموصــولة) : désignent une ou plusieurs personnes (animaux ou choses), comme par exemples الذي، التي، اللواتي، الذين ,etc.
13. conditionnel (اســم الشــرط) : ces mots sont considérés comme noms par les grammairiens arabes et qui ont une influence sur les verbes, en les mettant à l'inaccompli 'mdjzoum'. Comme par exemple : م، مـتىمهما، من، ا, etc.
14. interrogatif (اســم الاســتفهام) : ces mots se placent au début des phases et peuvent être précédés d'une patricule. Comme par exemple : كيــف، ماذا، لمـاذا, etc.
15. allusif (اســم الكنايــة) : ce sont quelques noms qui sont employés pour éviter de citer des expressions. Comme par exemple : كـذا، كـم , , etc.

16. circonstanciel (اسـم الضـرف) : c'est un nom de temps et de lieu qui précise quand, ou bien, où se réalise le verbe. Comme par exemple : الآن، بينمـا، حيـث, etc.
17. le nom de nombre (اسـم العـدد) : on distingue les noms de nombres cardinaux et les noms de nombres ordinaux. Comme par exemple : ثلاثـة، الاولـى، ألـف'

Bibliographie

[Abel, 04] Abel M.-H. (2004) "Utilisation de norms et standards dans le projet MEMORAe", In Distances et savoirs. Vol 2/4, 2004, pp. 487-511.

[Ahcene , 05] : Ahcene BENAYACHE. CONSTRUCTION D'UNE MEMOIRE ORGANISATIONNELLE DE FORMATION ET EVALUATION DANS UN CONTEXTE ELEARNING: LE PROJET MEMORAE. Thèse présentée pour l'obtention du grade de Docteur de l'UTC. décembre 2005.

[Aliane et al, 06]: ALIANE Hassina, ALIMAZIGHI Zaia, CHAOUI Aicha. Une approche basée Ontologie pour l'indexation automatique et la recherche d'information multilingue. Article accepté à la conférence IA'06 Maroc, 2006.

[Amandine, 05] : Amandine Schuurman. Recherche de services bioinformatiques dans une ontologie Investigation de la relation part-whole. Mémoire présenté en vue de l'obtention du grade de Maître en Informatique. 2005.

[Amsili, 03] Amsili, P. (2003). L'antonymie en terminologie : quelques remarques. In Actes des cinquièmes rencontres Terminologie et Intelligence Artificielle (TIA 2003).

[Bao-Quoc, 04] : HO Bao-Quoc.thèse de doctorat. Vers une indexation structurée basée sur des syntagmes nominaux (impact sur un SRI en vietnamien et la RI multilingue). Novembre 2004.

[Bateman et al, 95] : Bateman, J., Magnini, B.&Fabris,G.(1995), TheGen-eralized Upper Model Knowledge Base: Organiza-tionand Use, in N.Mars, ed., 'Towards Very Large Knowledge Bases', IOS Press.

[Baziz, 05] : Mustapha BAZIZ. Thèse pour l'obtention du doctorat. INDEXATION CONCEPTUELLE GUIDEE PAR ONTOLOGIE POUR LA RECHERCHE D'INFORMATION. Décembre 2005.

[Berners-Lee, 01]: Berners-Lee T., J., Lasilla, O., the semantic web, scientific American. 2001.

[Borst, 97] : Borst,P.(1997), Construction of Engineering Ontolo-gies for Knowledge Sharingand Reuse, PhDthesis,TweenteUniversity.

[Boughanem, 05] : cours poste graduation du module recherche d'information, Boughanem M. 2005.

[Boulaknadel, 06] : Siham Boulaknadel. Utilisation des syntagmes nominaux dans un système de recherche d'information en langue arabe. 2006.

[Bourigault, 96]: D. Bourigault, LEXTER, a Natural Language Processing Tool for Terminology Extraction, In Proceedings of 7th EURALEX International Congress, 1996.

[Bourigault, 98] : Bourigault, Condamines. Terminologies et ingénirie des connaissances. Bultin de l'AFIA, 1998, N°32. p 19-24.

[Bourigault et al., 05] Bourigault D., Fabre C., Frérot C., Jacques M.-P. et zdowska S. : Syntex, analyseur syntaxique de corpus. in Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles, France. 2005

[Brachman, 77] R.J. Brachman, What's in a concept : structured foundation for semantic networks, International Journal of Man-Machine Studies 9, pp 127-152, 1977.

[Carol, 01]: Carol Peter. Accès multilingue aux systèmes d'information. 67th IFLA Council and General Conference. August 16-25, 2001 http://www.ifla.org/IV/ifla67/papers/056-98f.pdf.

[CHEI, 92] : CHEI, M., MUGNIER, M-L.. Conceptual graphs : fundamental notion. Revue d'intelligence artificielle.1992. vol 6. p 275-279.

[Christian, 00] : Christian Jacquemin?, Pierre Zweigenbaumy. Traitement automatique des langues pour l'accès au contenu des documents. 2000.

[Concerto, 06]: The CONCERTO project, consulté le 24/06/06. http://www.apim.ens.fr/workshop_text_99/zarri_abstract.html.

[Christophe, 04] : CHARLES Christophe. thèse pour l'obtention du doctorat. SearchXQ : une méthode d'aide à la navigation fondée sur Ω-means, algorithme de classification non-supervisée. Application sur un corpus juridique Français. décembre 2004.

[Gruber et al , 94]: Gruber,T.&Olsen,G.R.(1994), An Ontology for Engineering Mathematics, in E.S.J.Doyle&P.Torasso,eds,'KR94Proceedings',MorganKauf-mann.

[Cruse, 00] Cruse, D. A. (2000). Meaning in Language. An Introduction to Semantics and Pragmatics. Oxford : University Press.

[Cunningham, 02] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, 2002.

[Daille, 96] Daille, B. (1996). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In Klavans, J. et Resnik, P., éditeurs : The Balancing Act : Combining Symbolic and Statistical Approaches to Language, pages 49–66. The MIT Press, Cambridge, Massachusetts.

[David, 90] S. David, P. Plante, Termino version 1.0, Report, Centre d'Analyse de Textes par Ordinateur, Université du Québec, 1990.

[Deerwester, 90]: S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer et R. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41, p. 391-407, 1990.

[Delafosse, 99] : Lionel Delafosse. Présentation du TAL. 1999.

[Delphine , 06] : Delphine BERNHARD. Thèse pour obtenir le grade de docteur. Apprentissage de connaissances morphologiques pour l'acquisition automatique de ressources lexicales, novembre 2006.

[Desmontils, 02] E. Desmontils, C. Jaquin, Indexing a Web site with a terminology oriented ontology, The Emerging Semantic Web, I. Cruz S. Decker, J. Euzenat, D.L. McGuinness (Eds.), IOS Press, ISBN 1-58603-255-0, pp 181-197, 2002.

[Domingue, 99] Domingue J. Motta E. (1999), A knowledge based news Server Supporting Ontology-Driven Story Enrichment and Knowledge retrieval. In Proceedings of EKAW'99 11th European Workshop on Knowledge Acquisition, Modelling and Management. Berlin: Springer Verlag, LNAI.

[Dumais, 03] S. Dumais, E. Cutrel, J. Cadiz, G. Jancke, R. Sarin, D. Robbins, Stuff I've Seen: A system for personal information retrieval and re-use, In Proceedings of the 26th ACM Conference on Research and Development in Information Retrieval (SIGIR'03), 2003.

[Elaine, 04] : Elaine Ménard. La recherche d'information multilingue. 2004. http://www.esi.umontreal.ca/~p0336101/RIML/indexa.html. date de consultation 01/05/2006.

[Enguehard et al, 92] : C. Enguehard - P. Malvache- P. Trigano. Indexation de textes : l'apprentissage des concepts. 1992.

[Enguehard, 92] Enguehard, C. (1992). ANA, Apprentissage Naturel Automatique d'un Réseau Sémantique. Thèse de doctorat, Université de Technologie de Compiègne.

[Fensel, 01] Fensel D. (2001)Ontologies: a silver bullet for Knowledge Management and Electronic Commerce. Berlin : Springer Verlag. 2001. Ce livre donne des définitions sur les ontologies et leurs application en industries et énumère les différents langages d'ontologie ainsi que les principaux sites et références traitant de ce sujet.

[Fluhr, 00] : C. Fluhr. Indexation et recherche d'information textuelle, Ingénierie des langues, Hermes. 2000.

[Gandon, 02] Fabien GANDON, « Ontology Engineering: a Survey and a Return on Experience », rapport de recherche INRIA, Mars 2002. Ce rapport présente l'objet "ontologie" et donne un état de l'art des techniques d'ingénierie d'ontologie. Puis présente le projet (CoMMA) pour lequel ils ont développé une ontologie. http://www.inria.fr/rrrt/rr-4396.html.

[Genest, 00]. David Genest. Thèse pour obtenir le grade de docteur. Extension du modèle des graphes conceptuels pour la recherche d'informations. 2000.

[Genest, 99] : David Genest. Vers un système de recherche documentaire basé sur les graphes conceptuels. Actes du XVIIème congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID'99). Consulté le 25/03/2008. http://134.214.81.35/articles/article169c1.pdf.

[Gomez, 99]: Gomez Pérez A., Benjamins V.R. (1999) "Overview of Knowledge Sharing and Reuse Components : Ontologies and problem-Solving Methods". Proceeding og the IJCAI-99 workshop on Ontologies and problem-Solving Methods (KRR5), Stockholm (Suède).

[Gruber, 93] T. Gruber, A translation approach to portable ontology speci_cation, Knowledge Acquisition, 7, 1993.

[Guha, 03] R.V. Guha, R. McCool, E. Miller, Semantic search, In Proceedings of the 12th International World Wide Web Conference, pp 700-709, 2003.

[Guiraude, 02] : Guiraude Lame.thèse pour obtenir le grade de docteur. Contruction d'ontologie a partir de textes : une ontologie du droit dédiée à la recherche d'information sur le web. Décembre 2002.

[Haddad, 02]. Mohamed Hatem HADDAD. Thèse pour obtenir le grade de docteur. Extraction et Impact des connaissances sur les performances des Systèmes de Recherche d'Information. Septembre 2002.

[Harbeck, 99]: S. Harbeck et U. Ohler. Multigrams for LanguageIdentification, Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech 99), Budapest, Hungary, 1999.

[Hayes, 1999] Hayes, B. (1999). The Web of Words. American Scientist, 87(2):108–112.

[IKSAL, 02] : IKSAL Sébastien. Thèse. Spécification Déclarative et Composition Sémantique pour des Documents Virtuels Personnalisables. Décembre 2002.

[Jac&Zwe, 00] : C. Jacquemin et P.Zweigenbaum. Traitement automatique des langues pour l'accès au contenu des documents. In J. Le Maître, J.Charlet and C. Garbay, editors, Le docu-ment Multimédia en Sciences du Traitement de l'Information, CةPADUبS-ةditions, Toulouse, p. 71-110, 2000.

[Jacquemin et Bourigault, 03] Jacquemin C. & Bourigault D. (2003). The Oxford Handbook of Computational Linguistics, chapter Term Extraction and Automatic Indexing, p. 599–615. Oxford Univesity Press.

[Kahan, 01] J. Kahan, M. Koivunen, E. Prud'Hommeaux, R. Swick, Annotea: An Open RDF Infrastructure for Shared Web Annotations, In Proceedings of the 10th International World Wide Web Conference, pp 623-632, 200.

[Khelif, 06] : Mohamed Khaled KHELIF. Thèse pour obtenir le grade de docteur. Web sémantique et mémoire d'expériences pour l'analyse du transcriptome. avril 2006.

[L'Homme, 01] : Marie-Claude L'Homme. Nouvelles technologies et recherche terminologique Techniques d'extraction des données terminologiques et leur impact sur le travail du terminographe. 2001.

[Lafourcade et Prince, 01] Lafourcade, M. et Prince, V. (2001). Synonymies et vecteurs conceptuels. In Actes de TALN 2001, pages 233–242, Tours, France.

[Lame, 02] G. Lame, Construction d'ontologie à partir de texte, une ontologie du droit dédiée à la recherche d'information sur le Web, Thèse de doctorat, Ecole des Mines de Paris, 2002.

[Laporte, 00] : E. Laporte. Mots et niveau lexical, Ingénierie des langues, Hermes, p. 25-49, 2000.

[Lebart et Salem, 94] : Lebart, L. et Salem, A. *Statistique textuelle*. Dunod, Paris. 1994

[Le Grand, 01] : Le Grand B. (2001) "Extraction d' information et visualisation de systèmes complexes sémantiquement structurés", thèse de doctorat de l'université de pierre et marie curie. décembre 2001.

[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*, sept. 89.

[Lounis, 06] : Louis Hébert. Le graphe sémantique. Consulté le 25/03/2008. http://209.85.129.104/search?q=cache:Cja-t5WQVLYJ:www.signosemio.com/rastier/graphe.asp+LE+GRAPHE+S%C3%89MANTIQUE%2BLouis+H%C3%A9bert&hl=fr&ct=clnk&cd=1&gl=fr. 2006.

[Malik, 02] : Mohamed Mahdi Malik. LE ROLE DE LA LOGIQUE FLOUE DANS LE WEB SEMANTIQUE. www.dil.univ-mrs.fr/dea/dea2002/memoires/malik.pdf. consulté le 24/06/06.

[Masolo, 01] Masolo C. (2001) Ontology driven Information retrieval: Stato dell'arte. Report of the IKF (Information and Knowledge Fusion) Eureka Project E!2235. LADSEBCnr, Padova (I).

[Mihalcea, 00] R. Mihalcea, D.I. Moldovan, Semantic Indexing using WordNet Senses, In Proceedings of ACL Workshop on IR & NLP, 2000.

[Miller, 90] : Miller,G.(1990), 'WordNet', InternationalJournalofLex-icography 3.

[Miller, 93] G. Miller, C. Leacock, R. Tengi, R.T. Bunker, A Semantic Concordance, In Proceedings of ARPA Workshop on Human Language Technology, pp 303-308, 1993.

[Mizoguchi, 00] : Mizoguchi R. et Bourdeau J. (2000) "Using Ontological Engineering to Overcome Common AI-ED Problems", In International Journal of Artificial Intelligence and Education, vol.11(Special Issue on AIED 2010).

[Morin, 98] Morin, E. (1998). Prométhée, un outil d'aide à l'acquisition de relations sémantiques entre termes. *In Actes de la conférence TALN 1998*,

[Namer, 03] Namer, F. (2003). Automatiser l'analyse morpho-sémantique non affixale : le système DériF. Cahiers de Grammaire.

[Nathalie, 05] : Nathalie HERNANDEZ. thèse pour obtenir le grade de docteur . ONTOLOGIES DE DOMAINE POUR LA MODELISATION DU CONTEXTE EN RECHERCHE D'INFORMATION. décembre 2005.

[On2broker, 04] : Ontobroker, http://www.ontoprise.de/products/ontobroker_en, consulté le 6 novembre 2004

[On2Knowledge, 02] : On-To-Knowledge : Content-driven Knowledge-Management through Evolving Ontologies. http://www.ontoknowledge.org/. Date consultation: 24/06/06.

[OntoSeek, 04] : Informations sur OntoSeek, consulté le 7 novembre 2004. http://babage.dia.fi.upm.es/ontoweb/wp1/OntoRoadMap/show_app.jsp?app_name=OntoSeek.

[P. Lefèvre, 00] : Philippe LEFEVRE. *La Recherche d'informations*, Hermès. 2000

[PHAN, 05] : PHAN Quang Trung Tien. Ontologies et Web Services. Sous la direction de Professeur NGUYEN Hong Quang. juillet 2005.

[Pierra, 02] G. Pierra, Un modèle formel d'ontologie pour l'ingénierie, le commerce électronique et le Web sémantique: Le modèle de dictionnaire sémantique PLIB , Journées Scienti_ques WEB SEMANTIQUE, Paris, 10-11/10/2002.

[Rijsbergen, 79] C.J. Van. Rijsbergen. Information Retrieval. Butterworths, London, 1979.

[Rijsbergen, 86]: Rijsbergen, C.J. a new theorical framework for information retrieval. Processings of the 9th annual international ACM SIGIR conference on research and devlopment in information retrieval, Pisa, Italy. Septembre 1986. p 194-200.

[Rint, 99] : Réseau international de néologie et de terminologie Revue semestrielle coéditée par l'Agence de la francophonie et la Communauté française de Belgique, N° 19, décembre 1998 :juin 1999.

[Robertson, 76]: S. E. Robertson & K. Spark Jones, 1976 : « Relevance Weighting for Search Terms », Journal of The American Society for Information Science, Vol 27, N°3, pp 129-146.

[Rebeyrolle, 00] Rebeyrolle, J. (2000). Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes. *In Actes de la conférence IC'2000, Journées Francophones d'Ingénierie de la Connaissance, Toulouse, IRIT*

[Roussey, 01]: Roussey Cathrine. Thèse pour obtenir le grade de docteur. Une méthode d'indexation sémantique adapté aux corpus multilingues, décembre 2001.

[SALT, 90] : SALTON, G., BUCKELEY, C. Improving Retrieval Performance by Relevance Feedback. Journal of the American Society for Information Science, 1990. Vol 41, N°4.p288-297.

[Salton, 70] Gerard Salton - Automatic processing of foreign language document – Journal of the American Society for Information Science, May 1970.

[Schwab *et al.*, 05] Schwab, D., Lafourcade, M. et Prince, V. (2005). Extraction semisupervisée de couples d'antonymes grâce à leur morphologie. *In Actes de TALN 2005*,

[SHOE, 06]: Simple HTML Ontology Extension. Date consultation: 24/06/06 http://www.cs.umd.edu/projects/plus/SHOE/.

[Sidhom, 02] : *S a h b i S I D H O M.* Thèse pour obtenir le grade de docteur. Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'information : *de l'écrit vers la gestion des connaissances. 2002.*

[SM, 83] : G. Salton and McGill. Introduction to Modern Information Retrieval. Mc Graw-Hill, New York, 1983.

[Smeaton, 89] : A. F. Smeaton. Information retrieval and natural language processing. In proceedings of a conference jointly sponsored by ASLIB, University of York, page 2, Mars 1989.

[Sowa, 84] J.F. Sowa. Ouvrage. Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley Publishing Company, USA, 1984.

[Strzalkowski , 93] : T. Strzalkowski. Natural language processing in large scale text retrieval tsks. In text Retrieval Conference (TREC-1), page 137, 1993.

[Swartout et al, 97] : Swartout, B., Patil, R., Knight, K.&Russ, T.(1997), To-ward Distributed Useof Large-Scale Ontologies, in 'AAAI97SpringSymposiumSeries,workshoponOntologicalEngineering'.

[TA, 05] : TA Tuan Anh, thèse, web sémantique et réseaux sociaux-contruction d'une mémoire collective par recommandations manuelles et (re)présentations. 2005.

[TALN, 05] : Methode du TALN, Traitement Automatisé du Langage Naturel, notion de l'indexation automatique.
http://www.educagri.fr/renadoc/telecharg/utilitaires/Indexauto.rtf. 2005

[Michel, 04] : Michel Gagnon. Introduction au web sémantique. 2004.

[Uschold *et a,* 96] : Uschold, M.&Gruninger, M.(1996), 'Ontologies:Prin-ciples, Methods and Applications', KnowledgeEngi-neeringReview 11.

[Vallet, 05] D. Vallet, M. Fernández, P. Castells, An Ontology-Based Information Retrieval Model, In Proceedings of the 2nd European Semantic Web Conference, pp 455-470, 2005.

[Vergne, 05] Vergne, J. (2005). Une méthode indépendante des langues pour indexer les documents de l'internet par extraction de termes de structure contrôlée. In Actes de la Conférence Internationale sur le Document Électronique (CIDE 8), Beyrouth, Liban.

[Véronique, 05] : Véronique Malaisé. Thèse pour obtenir le grade de docteur. Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles `a partir de corpus textuels.

[VIGNAUX, 04] : Georges VIGNAUX. La recherche d'information *Panorama des questions et des recherches. 2004*.
http://plate-forme-ast.mshparisnord.org/IMG/pdf/La_recherche_d_info.pdf.

[WebKB, 06] : WebKB, http://meganesia.int.gu.edu.au/~phmartin/WebKB/, consulté le 24/06/06.