

INDEXATION SEMANTIQUE DE CORPUS MULTILINGUES : APPLICATION AUX MANUSCRITS ANCIENS

Catherine ROUSSEY, Sylvie CALABRETTO, Jean-Marie PINON.

Laboratoire d'Ingénierie des Systèmes d'Information - INSA de Lyon.

20, avenue Albert Einstein

F-69 621 Villeurbanne Cedex.

Tel : 04 72 43 84 81 / Fax : 04 72 43 85 18

E-mail : croussey@lisisun1.insa-lyon.fr, cala@if.insa-lyon.fr, pinon@if.insa-lyon.fr

Résumé

Notre étude s'inscrit dans un Projet Européen LIBRARIES d'aide à la consultation et au travail des manuscrits anciens nommé BAMBI (Better Access to Manuscripts and Browsing of Image). Nous proposons une méthode d'indexation adaptée aux corpus multilingues. Cette méthode intègre un modèle de représentation sémantique des documents issu de travaux récents de notre équipe. Le formalisme de représentation du sens est basé sur les graphes conceptuels. Ce modèle définit une structure hiérarchique d'éléments sémantiques. Cette structure contient des informations sur la rhétorique et le contenu sémantique des différentes traductions d'un même document. L'index unique pour toutes les versions d'un document est constitué de la structure sémantique et de ses instances, ce qui nous permet de retrouver un document sans tenir compte de la barrière de la langue. Pour valider notre méthode nous l'avons appliquée à la transcription d'un document médical datant du XVIII^{ème} siècle.

Mots-clés:

indexation, recherche d'information multilingue, structure sémantique, documentation structurée, manuscrits anciens, représentation des connaissances.

Abstract

Our study takes place in an European Project LIBRARIES named BAMBI (Better Access to Manuscripts and Browsing of Image). This paper presents a method for multilingual access information. Our approach uses a semantic structure in order to index a document without taking account of its language. This semantic structure is a hierarchical organisation of semantic elements. Each element contains two levels of description, rhetorical organisation and meaning representation. The meaning representation is based on a conceptual graph formalism. A semantic structure is used as an index for a document and its translations. We applied our method to a transcription of a medical manuscript dated from the eighteen century.

Keywords

indexing, multilingual information retrieval, semantic structuring, structured document, old manuscript, meaning representation.

1. Introduction

Dans cet article, nous proposons une méthode d'indexation sémantique pour les corpus multilingues, c'est-à-dire contenant des documents écrits dans différentes langues. Notre proposition intègre un modèle de représentation sémantique des documents issu des travaux de recherche de Line Poulet [POUL97]. Nous allons essayer de construire une structure sémantique adaptée aux documents d'un corpus multilingue, dont la représentation du sens est basée sur les graphes conceptuels [SOWA84]. La structure sémantique est une forme d'index représentant le contenu d'un document et de toutes ses versions traduites, elle contient des termes descripteurs, représentatifs d'un concept. Nos travaux s'effectuant dans le cadre d'un projet de recherche européen intitulé BAMBI¹ (Better Access to Manuscripts and Browsing of Image) [CALA97], nous avons choisi de valider notre méthode avec la transcription d'un manuscrit médical datant du XVIII^{ème} siècle. Ce manuscrit est écrit en français et en italien. En conclusion, après une discussion sur notre méthode d'indexation multilingue, nous présenterons les perspectives de ce projet.

2. Indexation multilingue

Le développement du "World Wide Web" dans différents pays soulève le problème de la recherche documentaire dans un corpus de texte multilingue. Par exemple, on peut se demander comment retrouver un document écrit en français à l'aide d'une requête écrite en anglais. Il s'agit d'un cas particulier de la recherche multilingue, plus communément appelé cross-language information retrieval (CLIR). Notre cadre de recherche se limite à ce cas, dont le thème est associé au problème plus général de la paraphrase : retrouver un document qui répond à une requête sans utiliser le même vocabulaire. Il ne faut plus rechercher des mots mais des concepts. Différentes approches de recherche multilingue ont été mises au point. Il existe trois grandes techniques basées respectivement sur la traduction automatique, le vocabulaire contrôlé et le corpus.

2.1 Technique basée sur la traduction automatique

La première idée développée a été de traduire la requête et les documents du corpus dans la même langue; puis d'effectuer une recherche documentaire monolingue. Par exemple, le système de traduction automatique SYSTRAN a été ajouté à SPIRIT [RADW91] pour traduire les requêtes. Malheureusement, la combinaison d'un traducteur automatique et d'un indexeur monolingue subit les inconvénients d'une mauvaise traduction. En effet, la traduction du terme français « *temps* » peut donner en anglais « *weather* » ou « *time* », le mauvais choix du terme anglais entraînera une inadéquation entre les concepts français et anglais et donc une erreur dans le processus d'indexation.

2.2 Technique basée sur un vocabulaire contrôlé

Il existe plusieurs façons de présenter le vocabulaire d'un langage documentaire multilingue, le thesaurus et le langage pivot.

¹ Les différents partenaires européens du projet BAMBI sont ACTA S.p.a. (Computer society of Florence), CNR (Consiglio Nazionale della Ricerche - Istituto di Linguistica Computazionale di Pisa), BNR (Biblioteca Nazionale Centrale V.E.II di Roma), MPI (Max Planck Institut für Rechtsgeschichte (München)), CPR (Consorzio Pisa Ricerche), et le LISI.

2.2.1 *Le thesaurus*

Un thesaurus contient un lexique de mots normalisés d'un langage documentaire. Ces termes sont reliés par des relations sémantiques (synonymes, termes associés, termes génériques...) qui illustrent les connaissances du domaine. Un thesaurus multilingue est construit en associant à la liste de termes choisis comme descripteurs, c'est-à-dire représentatifs d'un unique concept, tous leurs synonymes dans plusieurs langues.

Pour chaque document, on sélectionne des descripteurs en tenant compte des connaissances sémantiques contenues dans le thesaurus: liste de synonymes, de termes associés, hiérarchie des termes, phrases illustrant le contexte d'utilisation d'un mot. On utilise la technique d'expansion de requête pour interroger la base [SALT94, BALL97], c'est-à-dire qu'on ajoute à la requête tous les termes reliés à ses descripteurs, synonymes, termes associés ou génériques. L'exemple le plus probant est le projet EMIR utilisant le logiciel SPIRIT [RADW91]. La requête, écrite en langage naturel, subit un traitement linguistique préalable pour retrouver ses descripteurs. Puis, on leur associe toutes leurs traductions possibles, sans tenir compte du sens. Dans ces traductions de la requête, il y a forcément la bonne expression et ce sera celle qui existe dans la base. En résumé, le corpus sert de filtre de traduction.

Ces techniques partent de l'hypothèse que les concepts sont indépendants de la langue, hypothèse non vérifiée dans certains cas. Par exemple, les anglais font une distinction entre la viande de mouton et cet animal vivant: "*mutton*" et "*sheep*" alors que les deux concepts sont réunis en français sous un même mot: "*mouton*". Plutôt que d'accumuler les connaissances dans diverses langues pour arriver à une indexation multilingue, une autre méthode est de synthétiser toutes ces connaissances puis de créer un langage pivot.

2.2.2 *Le langage pivot.*

La méthode basée sur le langage pivot consiste à identifier toutes les connaissances possibles d'un domaine et à les exprimer dans un langage appelé "langage pivot", indépendant des langues du corpus. Il contient non seulement les concepts du domaine mais aussi des relations. La recherche documentaire s'effectue en traduisant le corpus et les requêtes dans le langage pivot qui sert de base de référence pour comparer les connaissances contenues dans tout document.

Par exemple, le système RECIT [RASS94] représente les connaissances d'un domaine particulier de la médecine sous forme des graphes conceptuels. L'ensemble des concepts et des relations définis dans les graphes constitue le langage pivot. Avant de pouvoir être interrogé, le contenu du document est transformé en un graphe conceptuel.

L'inconvénient majeur du vocabulaire contrôlé (thesaurus et langage pivot) est qu'il n'intègre pas les mots nouveaux. En effet, si un terme n'est pas reconnu, c'est à dire s'il n'appartient pas au vocabulaire, il ne sera pas pris en compte dans l'indexation. Par conséquent, il n'aura aucun impact dans la recherche documentaire.

2.3 **Techniques basées sur le corpus**

Cette approche s'appuie sur un corpus parallèle, où chaque document existe en plusieurs versions, l'original et ses traductions. Ce corpus particulier est analysé à l'aide de techniques statistiques [DUMA90, SALT94, DAVI97] représentant les documents dans un espace de concepts. Ces techniques travaillent sur la répartition des termes dans les paragraphes pour regrouper les mots ayant la même signification, désignant le même concept.

Le corpus utilisé est composé de trois parties: une partie de documents français, une partie de documents anglais et un corpus parallèle contenant des documents français plus leurs versions anglaises et inversement.

Seul le corpus parallèle est interrogé par la requête. Les documents de chaque langue, retrouvés comme les plus pertinents, servent de requête pour interroger le reste de la base [FLUH95].

Bien que ces méthodes n'excluent pas le vocabulaire nouveau, elles sont très limitées. En effet, il existe peu de corpus parallèles.

Les domaines de recherches actuels se concentrent sur la création automatique de thesaurus multilingue à l'aide d'un corpus parallèle. Les deux approches basées sur le thesaurus et sur un corpus se joignent dans cette nouvelle perspective, sans résoudre d'ailleurs le problème de la rareté des corpus parallèles. Pour le moment aucune approche ne donne entièrement satisfaction pour résoudre les ambiguïtés de sens des groupes de mots complexes, c'est à dire pour déterminer correctement la sémantique d'une phrase. Chacune de ses approches cherche à traduire les termes de la requête dans le langage du document, donc nous rejoignons les problèmes de traduction. Par conséquent, nous avons choisi de nous intéresser particulièrement à la sémantique des textes, en essayant de séparer le sens du document de sa langue d'écriture. Nous sommes donc partis de l'approche du langage pivot et nous l'avons amélioré en intégrant différents niveaux de connaissances pour permettre une indexation plus riche donc une recherche documentaire plus efficace. Notre approche intègre un langage pivot dans une structure sémantique, pour lui permettre d'exprimer plusieurs niveaux de connaissances, la rhétorique et le contenu sémantique.

3. Proposition d'une méthode d'indexation sémantique multilingue.

3.1 Modèle de représentation de document

D'après de récentes recherches [NANA96], un document peut être lu de manières différentes, suivant le point de vue du lecteur. Par exemple, l'imprimeur se concentrera sur la représentation physique du document; l'auteur s'intéressera à l'organisation du texte et le simple lecteur cherchera le sens du document; par conséquent, un document contient au moins trois structures de représentation.

La structure physique donne la présentation du document sur le papier; elle est constituée d'éléments physiques tel que la première page, un cadre, une colonne.

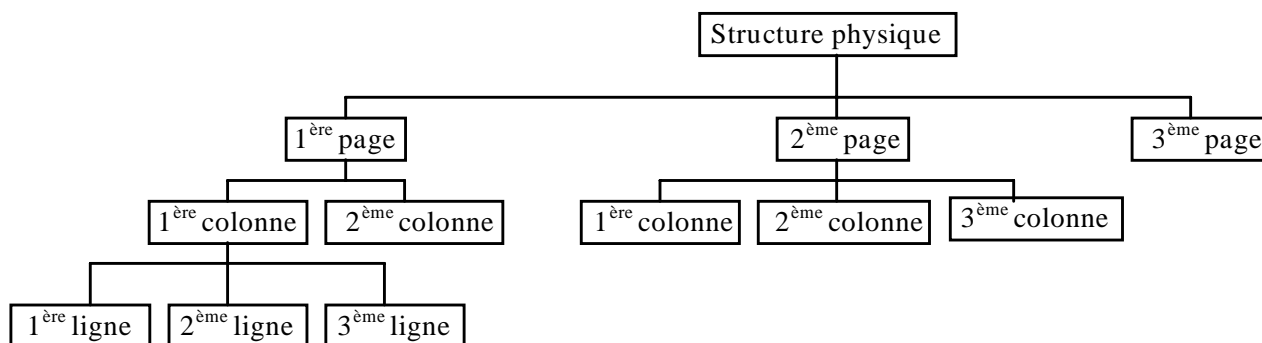


Figure 1 : Exemple de structure physique.

La **structure logique**² retrace l'organisation de l'information contenue dans le document. Les éléments logiques la composant sont les titres, les chapitres, les paragraphes, les notes, les schémas.

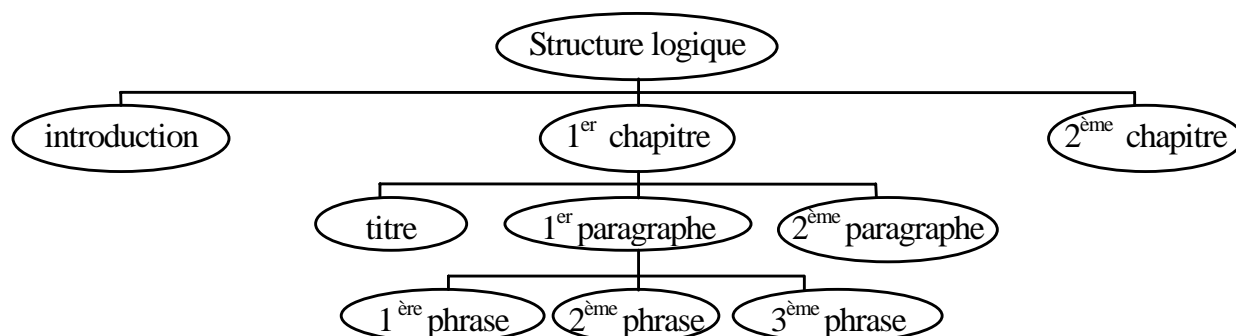


Figure 2 : Exemple de structure logique.

La **structure sémantique** représente l'information en elle-même, le sens du contenu du document [NANA96], [POUL97]. Un élément sémantique représente une relation entre concepts. L'information pouvant être décrite de diverses manières, nous avons choisi un formalisme tiré des graphes conceptuels [SOWA84] pour représenter le sens d'une phrase; ce modèle sera approfondi plus loin [cf. § 3.2].

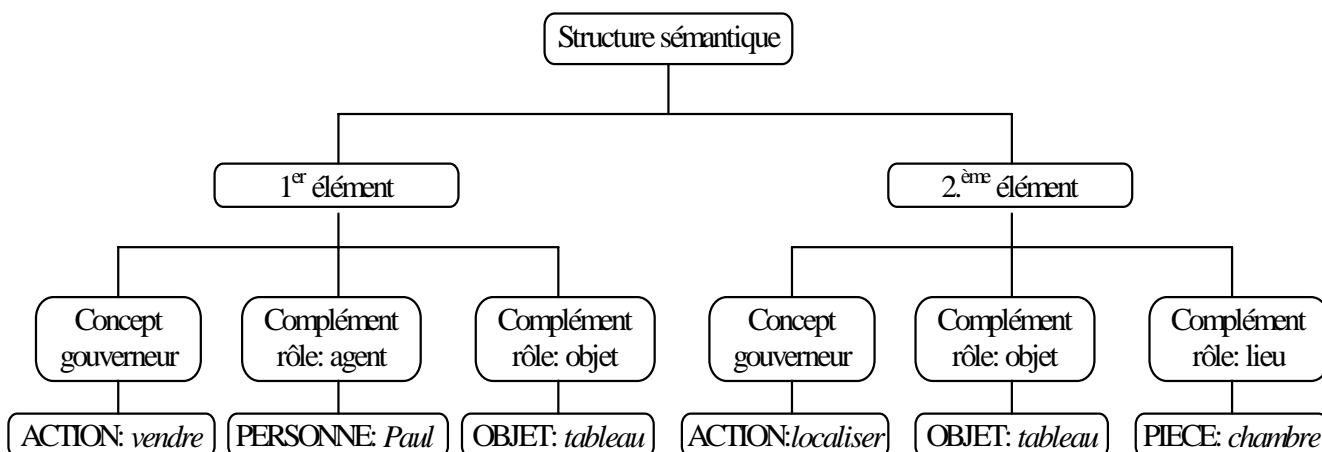


Figure 3 : Structure sémantique de la phrase

"Paul vend un tableau; celui qui se trouve dans la chambre".

Ces trois représentations sont structurées hiérarchiquement par une arborescence de leurs éléments caractéristiques physiques, logiques et sémantiques. Ces éléments contiennent une séquence d'objets élémentaires ou composés. Chacun des objets élémentaires, les feuilles terminales de l'arborescence, se rattache à une portion de contenu du document initial, qui correspond à une instance d'un objet élémentaire. Des caractéristiques typiques de la structure sont attribuées à cette portion de document. En résumé, une portion de contenu est décrite trois fois de manière différente, suivant la structure dans laquelle on se place (physique, logique ou sémantique).

² Nos exemples de structures physique et logique ne correspondent pas aux structurations les plus courantes. Effectivement, les éléments terminaux ne sont généralement pas associés à des objets aussi précis que la ligne physique et la phrase logique; c'est pour des raisons de compréhension que nous avons choisi d'affiner nos exemples.

Article V Deux Doses Fébrifuges
Prenez une once de Kinkina en poudre, une livre de bon vin vieux. Laissez les ensemble pendant quatre heures, ensuite ajoutez-y deux livres d'eau bouillante; laissez-les encore pendant six heures au moins dans un vaisseau de verre. Versez les doses à clair, vu (en terme de pharmacie) par inclinaison quand on voudra s'en servir. Chaque dose sera de six onces.

p13

Figure 4 : Exemple de recette extrait « Des nouvelles formules de médecines ».

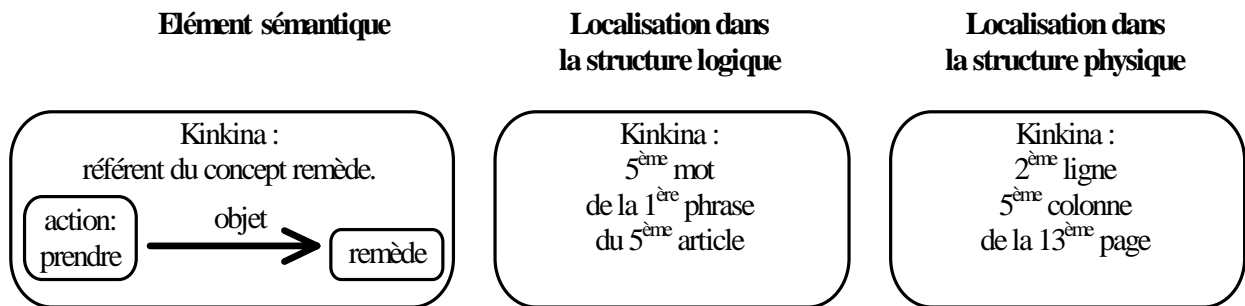


Figure 5 : Exemple de localisation d'un élément du document dans les trois structures.

Ces trois structures existent à part entière individuellement mais des liens permettent de passer d'une structure à une autre. Ces structures sont associées au même document. Nous pouvons citer trois types de liens entre les différentes structures:

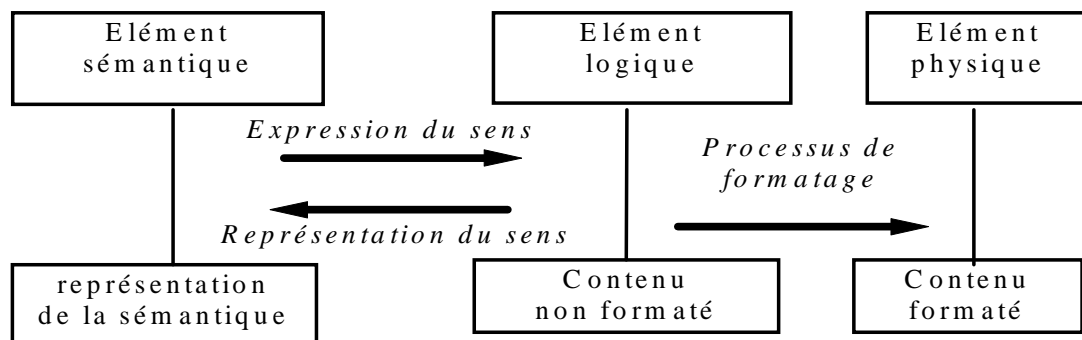


Figure 6 : Liens entre les modèles.

Les liens bidirectionnels entre la structure sémantique et la structure logique nous permettent de détacher la structure sémantique du contenu documentaire. Pour retrouver la portion de document rattaché à un élément sémantique, il faut passer par la structure logique. Par ce biais, l'expression du sens devient indépendante de la langue du document. Nous partons de l'hypothèse qu'à un bas niveau de granularité, le sens d'un mot³ peut être considéré comme indépendant de la langue. Par contre, les éléments de la structure logique eux restent liés à la langue d'écriture du document parce qu'ils sont instanciés par les portions correspondantes du document. En conséquence, partant d'un concept défini dans une structure sémantique on peut retrouver ses synonymes dans une langue, par la structure logique du document. Les documents seront indexés à partir de leur structure sémantique et ils seront restitués à l'aide des liens

³ Nous savons que cette hypothèse est fautive dans un contexte général. Cependant, pour indexer un document nous sommes amenés à simplifier le sens d'une portion de document. Notre approche de l'indexation représente des connaissances généralistes [NANA96] et non des connaissances expertes. (cf. § 2.2.1).

d'expression du sens. Notre travail s'intéresse donc uniquement aux relations entre les structures logique et sémantique.

3.2 Proposition d'un modèle sémantique pour la représentation des connaissances multilingues

Notre proposition consiste à construire une structure sémantique adaptée aux corpus multilingues. L'idée est de définir une structure pour organiser de manière rigoureuse l'information contenue dans le document [POUL97A,B]. Cette information sera formulée dans un langage indépendant des langues du corpus étudié. Cette structure permettra, entre autre, d'interroger le corpus de documents comme une base de données (l'ensemble des concepts). De même que les structures logiques et physiques, il s'agit d'une arborescence d'éléments sémantiques dont les feuilles terminales sont des objets élémentaires :

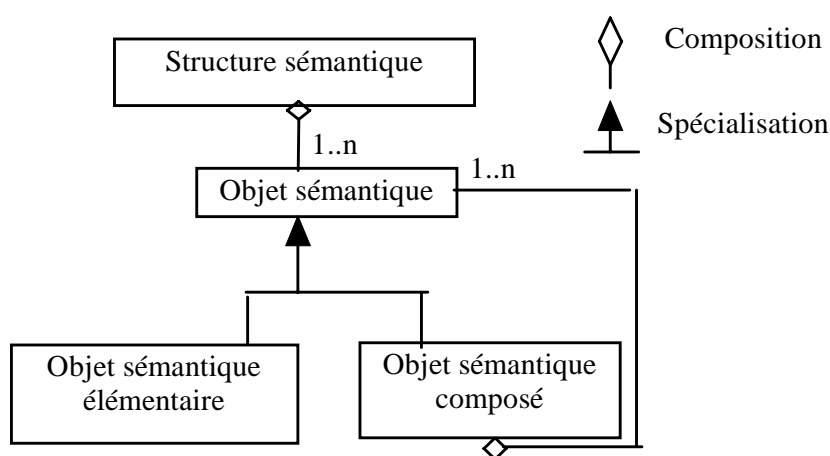


Figure 7 : La composition de la structure sémantique

La structure sémantique reflète l'organisation du discours dans le document par une hiérarchie d'éléments sémantiques composés, caractérisant les types de discours ou des relations entre objets. Les objets sémantiques sont typés pour identifier la rhétorique du discours (descriptions, observations) [NANA96]. Plus précisément, les objets sémantiques formalisent le sens des contenus documentaires. La composition des objets élémentaires est une représentation de la sémantique des langues naturelles basée sur les graphes conceptuels.

3.2.1 Brève description du formalisme de représentation des connaissances

La sémantique du contenu documentaire est exprimée par un formalisme basé sur les graphes conceptuels (GCs) [SOWA84] et sur la logique terminologique issue du langage KL-ONE [BRAC85]. Notre formalisme a pour objectif de regrouper les points communs de ces deux méthodes de représentation des connaissances en cumulant leurs atouts, c'est-à-dire la richesse d'expressivité des GCs et la modélisation pragmatique proche des langages informatiques de la logique terminologique. La recherche d'information se fait sur la base des travaux de Sowa qui établit grâce à l'opérateur Φ une connexion entre les GCs et le modèle logique de recherche d'information [RIJS86]. Cet opérateur traduit les relations entre graphes par des expressions de la logique du premier ordre.

Les connaissances sont représentées par des liens (appelés relation ou rôle) entre des noeuds (appelés concepts). Les concepts sont des entités atomiques identifiées par un **label de type**. Un concept peut être instancié par un *réfèrent*, par exemple un objet du monde réel. Les relations spécifient les rapport entre les concepts, elles sont aussi caractérisées par un *type*. Dans nos GCs, les liens partent tous d'un concept gouverneur vers des concepts ordinaires [OUNI95].

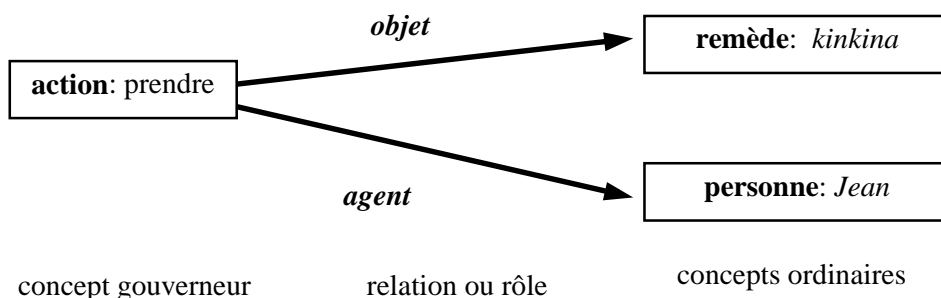


Figure 8: Exemple de Graphe Conceptuel.

Une relation de généralisation appelée relation de subsumption est établie entre les types de concepts. En d'autres termes un concept de type A subsume un concept de type B si A est plus général que B, toutes les instances (ou référents) de B sont aussi des instances de A. L'ensemble de la hiérarchie est organisée dans un treillis, ayant pour sommet le type universel, et pour base le type absurde.

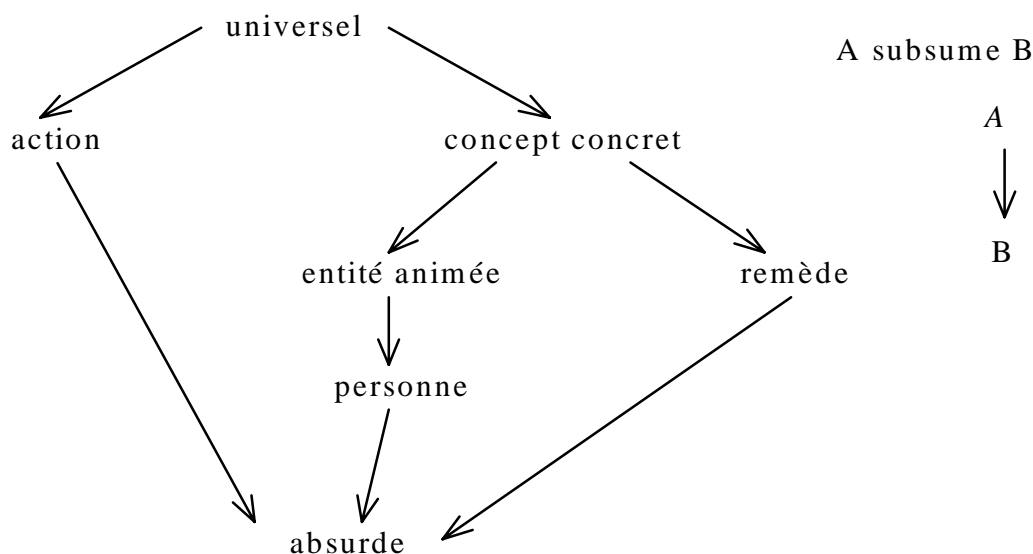


Figure 9 : Treillis de concept.

3.2.2 Objet sémantique élémentaire

Un objet sémantique élémentaire représente un fragment d'information. Il existe deux niveaux de connaissance:

- La rhétorique définie par des types de discours. Ceux-ci forment un métalangage caractérisant le rôle que joue un fragment de texte dans la transmission de l'information à l'homme. Ces types sont par exemple: définition, exemple, argumentation, spécification [NANA96].

- Le contenu sémantique des fragments d'information exprimé par des relations entre concepts [Figure n° 8]. C'est à ce niveau qu'apparaissent les connaissances du domaine.

La structure d'un objet élémentaire représente la sémantique d'un fragment d'information. Tout objet sémantique peut être typé pour exprimer la rhétorique. Les instances d'un objet sémantique sont les référents des concepts de l'objet.

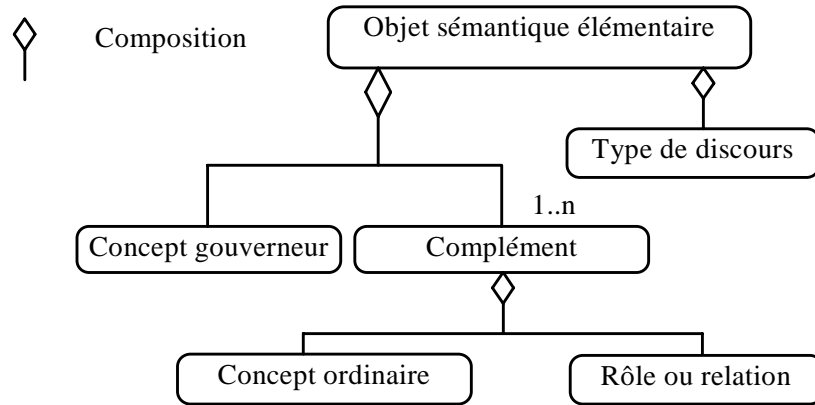


Figure 10 : Composition d'un objet sémantique élémentaire.

3.3 Lien entre le modèle sémantique et le modèle logique

En partant du principe que le sens est indépendant de la langue, un document écrit en français et sa traduction en italien fournissent les mêmes informations. Par conséquent, les deux versions de ce document ont la même structure sémantique avec le même contenu. Les éléments de cette structure ainsi que leurs instances sont écrits dans un langage pivot. Le langage pivot exprime entre autre les rôles sémantiques, les concepts et leurs référents. Nous avons choisi à titre d'exemple comme langage pivot la langue de communication scientifique, l'anglais. La figure n°11 nous présente un exemple de lien entre la structure sémantique et la structure logique. L'instance de l'élément sémantique contient *to take*, référent du concept **Action**, *quinquina et wine*, référents du concept **Drug**.

Au contraire, les instances des éléments logiques, par exemple un titre de livre « Les Remèdes Correctifs », dépendent de la langue du document. Donc, en établissant des liens entre des éléments logiques et sémantiques (ou leur instances respectives) se rapportant à la même portion de document, nous obtenons des liens de représentation du sens ou plus précisément de traduction du concept dans diverses langues.

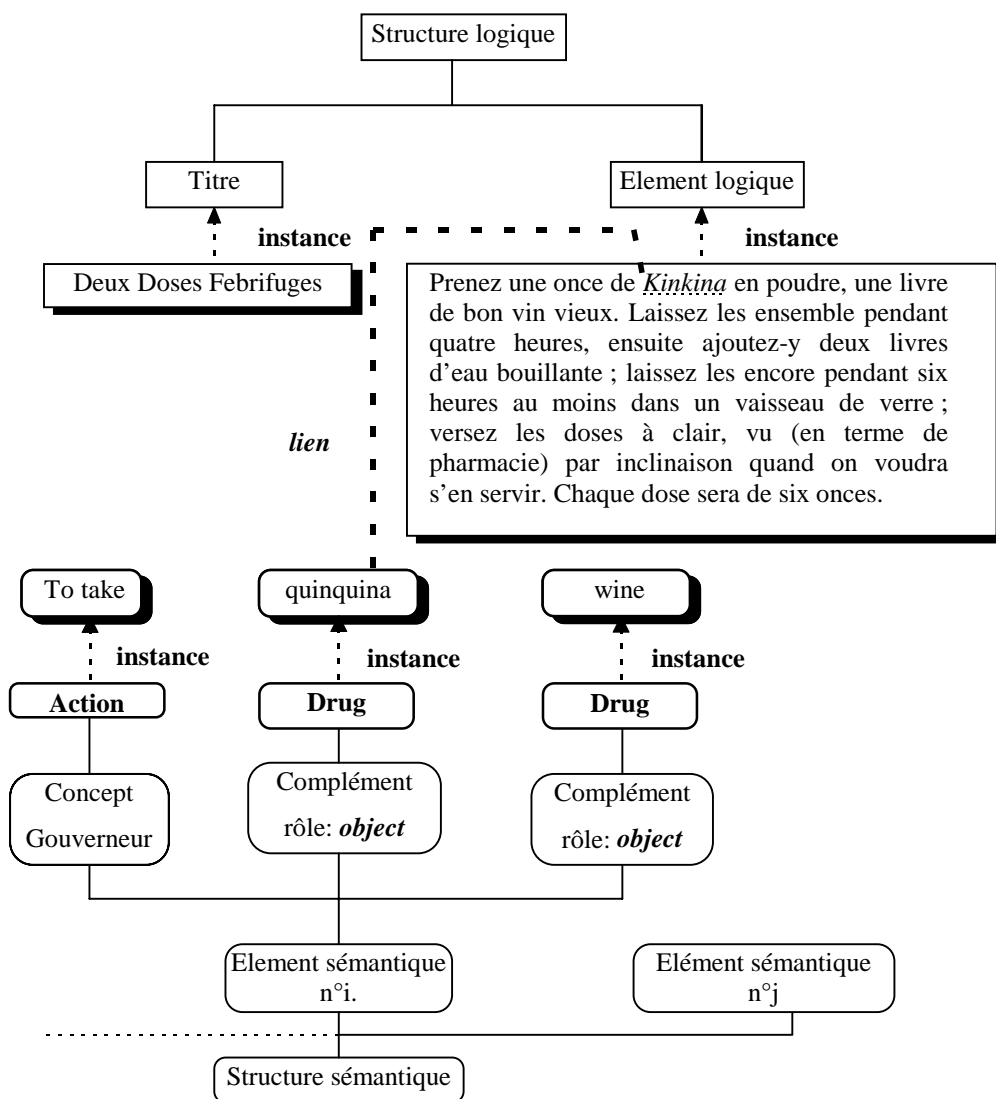


Figure 11 : Exemple de lien de traduction entre les instances sémantique et logique.

4. Application : un manuscrit médical du XVIII^e S.

Notre recherche s'inscrit dans un projet de Recherche Européen LIBRARIES intitulé BAMBI⁴ (Better Access to Manuscripts and Browsing of Image) [CALA97]. Ce projet vise à concevoir un système interactif pour la consultation et le travail sur des manuscrits anciens. Les manuscrits sont des documents écrits à la main, souvent intéressants du point de vue graphisme (exemple: les enluminures). Le système BAMBI met à la disposition de l'utilisateur, des images du manuscrits ainsi qu'un fichier texte contenant sa transcription. La transcription, réalisée par un philologue, est un processus visant à "traduire", expliciter, les nombreuses abréviations contenues dans les textes anciens (par exemple, bul pour Bulgarius, ac pour Accarcus). Le corpus multilingue qui nous intéresse est constitué de toutes les transcriptions disponibles dans BAMBI. En effet, pour le moment, la recherche d'un manuscrit se fait par l'intermédiaire d'informations signalétiques,

⁴ Les différents partenaires européens du projet BAMBI sont ACTA S.p.a. (Computer society of Florence), CNR (Consiglio Nazionale della Ricerca - Istituto di Linguistica Computazionale di Pisa), BNR (Biblioteca Nazionale Centrale V.E.II di Roma), MPI (Max Planck Institut für Rechtsgeschichte (München)), CPR (Consorzio Pisa Ricerche), et le LISI.

c'est-à-dire le titre, la langue, la date, les auteurs, etc.... Notre objectif vise à étendre cette recherche à la sémantique des transcriptions.

A l'aide d'un exemple de manuscrit nous allons détailler comment on peut relier un concept appartenant à la structure sémantique à ses différentes traductions par le biais de la structure logique. Ce manuscrit daté de 1716 est un livre de pharmacologie contenant les recettes pour la fabrication de différents remèdes. Intitulé « *Les nouvelles formules de médecines* », il a été écrit par un docteur en médecine, Pierre Garnier. Chaque recette est écrite en français et en italien. Sa structure logique se décompose de la manière suivante :

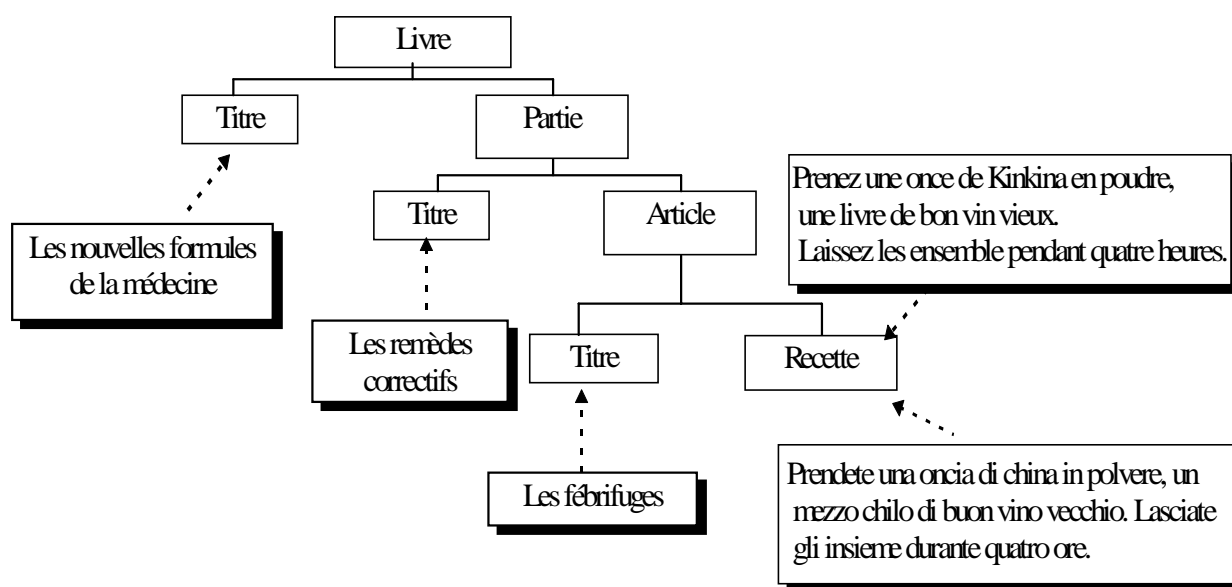


Figure 12 : Structure logique « *Des nouvelles formules de médecines* ».

Dans ce cas précis, la rhétorique du discours est simple, car les recettes sont une succession de descriptions d'étapes. Une étape est constitué en général d'une action, le concept gouverneur (*prendre, laisser reposer*) et de plusieurs autres concepts. Chacun est relié au concept gouverneur par un rôle définissant la relation existant entre le concept gouverneur et le concept ordinaire (*objet* subissant l'action). Par exemple, la première phrase de notre recette peut être décrite par le graphe suivant [figure n°13]. Nous disposons donc d'un graphe complexe regroupant plusieurs sous graphes ayant chacun leur concept gouverneur.

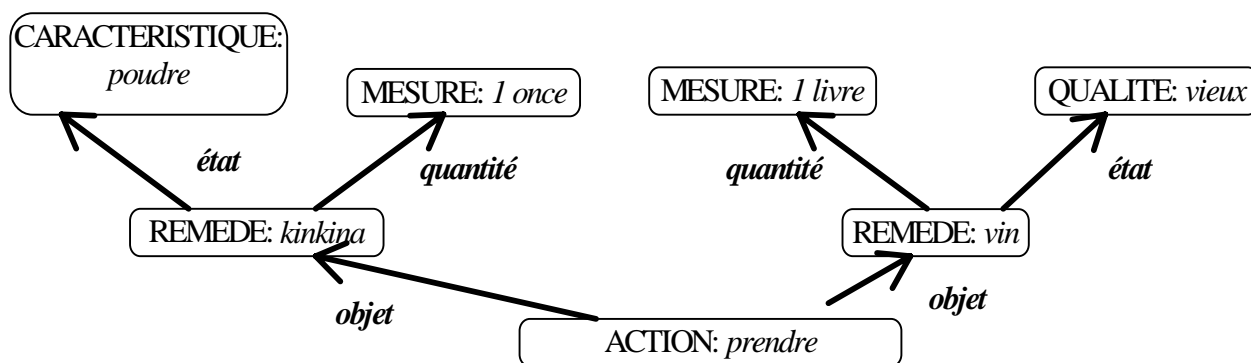


Figure 13 : Exemple de graphe conceptuel.

Les rectangles représentent des concepts (action, remède) et leur référents (*kinkina*, *vin*, *laisser ensemble*). Les arcs indiquent une relation. Ce graphe est décrit dans un élément sémantique étape. Une partie de la composition de cet élément est détaillée dans la structure sémantique du manuscrit.

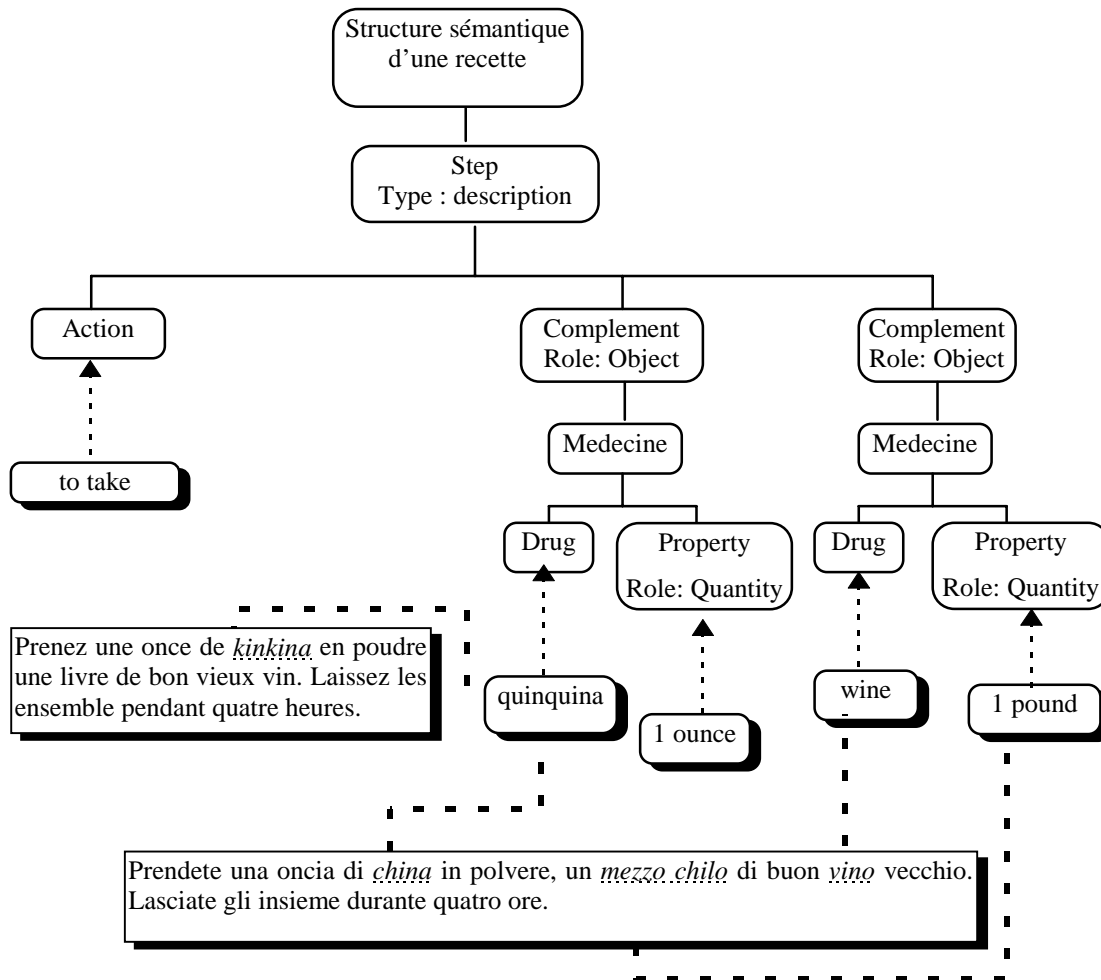


Figure 14 : Exemple de lien de traduction «Des nouvelles formules de médecine».

Cet exemple montre que les instances des éléments sémantiques peuvent être reliées à plusieurs instances d'une même classe d'élément logique. Ce qui permet de traduire en plusieurs langues un concept contenu dans un élément sémantique.

Dans notre exemple de structure sémantique nous avons un concept dont le label type est **Drug**. C'est un élément sémantique désignant une connaissance du domaine de la pharmacologie. On peut remarquer que, dans ce contexte particulier, le vin est aussi un remède.

4.1 Recherche d'information dans les corpus multilingues à partir de notre méthode.

Comme nous l'avons vu, la structure sémantique contient plusieurs niveaux de connaissances (la rhétorique et le contenu sémantique). La recherche documentaire va utiliser ces différents niveaux pour retrouver des documents.

A l'aide des éléments sémantiques représentatifs des connaissances du domaine, on peut construire des catalogues et donc interroger la structure comme une base de connaissances. Par exemple, les instances de l'élément **Drug** (*wine, quinquina*) forment un catalogue de tous les remèdes existants dans l'ensemble du corpus. En parcourant ce catalogue, on peut choisir quel remède nous intéresse et retrouver tous les documents correspondants, quelle que soit leur langue. La construction de catalogues de noms de plantes, de noms de personnes, est très importante en philologie, car la première tâche d'un philologue est de répertorier ce type de connaissances [ANDR94].

On peut aussi interroger cette structure à l'aide d'une requête en langage pivot. Cette requête se formalise comme un objet sémantique élémentaire avec l'ensemble des instances de ses concepts. Donc on pourra associer le niveau rhétorique au niveau sémantique pour interroger notre corpus. En effet, le langage pivot n'est pas constitué uniquement des instances sémantiques mais aussi des éléments sémantiques avec leur type rhétorique et leur rôle. L'index d'un document, c'est à dire les connaissances impliquées dans un document et qui permettent de le retrouver, n'est plus uniquement un ensemble de descripteurs, par exemple l'ensemble des instances sémantiques, mais la structure sémantique dans son intégralité.

A un niveau plus général, on peut parcourir les structures sémantiques pour faire une recherche par thème. Il s'agit d'une recherche assistée où l'utilisateur est guidé, car il n'a qu'un nombre limité de choix possibles. Effectivement, cette structure est un arbre qui part d'un thème très général (la pharmacologie) et qui se termine par des éléments au sens beaucoup plus précis (la fabrication d'un médicament).

Pour pouvoir construire une structure sémantique, il faut répertorier les connaissances, qu'elles soient propres au domaine ou plus générales, c'est-à-dire l'ensemble des concepts, des référents, et de leur relation, (leur rôle possible vis à vis d'un concept gouverneur). Ces connaissances vont permettre de construire le langage pivot. Plus précisément, pour chaque concept, il faudra choisir un seul et unique terme, son descripteur. De plus, il faut bien connaître les besoins des utilisateurs. Non seulement, il faut savoir quel genre de connaissances nécessite la construction de catalogue (plante, lieu, personne), mais aussi ce que l'utilisateur juge important sémantiquement. C'est cette vue du document qui va influencer l'organisation des éléments sémantiques. Naturellement, la composition de la structure dépend de ce que l'on désire représenter. Nous débouchons sur le paradigme de la représentation des connaissances [MART95]. En ce qui nous concerne, nous avons favorisé la représentation de certaines connaissances utiles aux philologues, leur travail consiste surtout à répertorier tous les noms propres.

5. Conclusion et perspectives

Dans ces quelques pages, nous avons essayé de décrire une méthode d'indexation multilingue basée sur la représentation sémantique des documents. L'index est devenu une structure sémantique avec ses instances. Cette structure contient deux niveaux de connaissance: la rhétorique, et le contenu sémantique, exprimée à l'aide d'un formalisme proche des graphes conceptuels. Ce type d'indexation permet d'interroger le corpus de différentes manières adaptées à différents types d'utilisateur. En effet, si celui-ci ne connaît pas le contenu du corpus, il peut bénéficier d'une recherche guidée en parcourant les structures sémantiques. Par contre, si l'utilisateur connaît bien le corpus, il peut soit formuler une requête en langage pivot, soit parcourir les catalogues, contenant l'ensemble des référents d'un concept du domaine, pour trouver ce qui l'intéresse.

Pour valider notre approche, nous avons défini des structures sémantiques avec le formalisme d'HyTime [NEWC91]. Ce qui nous a permis de créer une structure sémantique générique, indiquant toutes les connaissances pouvant être utilisées, connaissances du domaine en particulier, et une structure sémantique spécifique exprimant les connaissances tiré d'un document et de ses traductions.

Même s'il nous paraît difficile d'automatiser totalement cette méthode d'indexation, il pourrait être envisageable de guider l'utilisateur dans la construction des structures sémantiques, en commençant par mettre à sa disposition toutes les connaissances que le système possède déjà (la hiérarchie des concepts et la liste de leur rôle possible). De plus, certains documents sont proches sémantiquement car ils appartiennent au même domaine. Il pourrait donc être intéressant de se baser sur des structures existantes pour en construire de nouvelles et ainsi réutiliser les mêmes éléments sémantiques. Pourquoi ne pas envisager dans l'absolu un processus d'apprentissage de composition de structures sémantiques ? Cet apprentissage capitaliserait les actions déjà réalisées par les utilisateurs pour les réutiliser et permettrait ainsi d'éviter la répétition des mêmes opérations.

6. Bibliographie

- [ANDR94] Jacques Andre, Hélène Richey. Utilisation des index d'un éditeur structuré dans le cadre d'actes médiévaux. Intelligence artificielle, systèmes cognitifs et interaction homme-machine. Projet Opéra. Publication interne de l'IRISA n°841, 50p, 27 juin 1994.
- [BALL97] Lisa Ballesteros, Bruce Croft. Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia PA, USA, p84-91, 1997.
- [BRAC85] R.J. Brachman, J.G. Schmolze. An overview of the KL-ONE knowledge representation systems. Cognitive science Vol.9(2), p171-216, 1985.
- [CALA97] Sylvie Calabretto, Jean-Marie Pinon. Modelling a Medieval Manuscript Database with HyTime. EP'97, Canterbury, à paraître dans Chapman & Hall, Avril 1997.
- [DAVI97] Mark Davis, William Ogden. QUILT: Implementing a Large-Scale Cross-Language Text Retrieval System. Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia PA, USA. p92-98, 1997.
- [DUMA90] Susan T. Dumais, George W. Furnas, Thomas K. Landauer. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, Vol 41, N° 6, p391-407, 1990.
- [FLUH95] Christian Fluhr. Multilingual Information Retrieval. Survey of the State of the Art in Human Language Technology, Center for Spoken Language Understanding, Oregon Graduate Institute, p 305-391, 1995.
- [MART95] P. Martin. Links between Electronic Documents and a Knowledge Base of Conceptual Graphs. In : Proc. of International Conf. of Conceptual Structures, California : University of California, p. 112-125, Août 1995.

- [NANA96] Marc Nanard, Jocelyne Nanard, Jacques Chauche, & Al. La métaphore du généraliste : Acquisition et utilisation de connaissances macroscopiques sur une base de documents techniques. Acquisition et Ingénierie des Connaissances - Tendances actuelles. N. Aussenac-Gilles, P. Laublet, C. Reynaud. Toulouse : CEPADUES, p 285-304, 1996.
- [NEWC91] S. Newcomb, N.A. Kipp, V.T. Newcomb. The HyTime Hypermedia/Time based document structuring language. Communication of the ACM, Vol.34, N°11, p.67-83, Novembre 1991.
- [OUNI95] Iadh Ounis. Une Dénotation pour les Graphes Conceptuels: comparaison avec les logiques Terminologiques en Recherche d'Information. Actes du XIII^e Congrès INFORSID, Grenoble, p.147-164, Juin 1995.
- [POUL97A] Line Poulet. Formaliser la sémantique des documents - Un modèle unificateur. XV^e Congrès INFORSID, Toulouse, p.339-352, Juin 1997.
- [POUL97B] Line Poulet, Jean-Marie Pinon, Sylvie Calabretto. Semantic Structuring Of Documents. Proceedings of the Third Basque International Workshop on Information Technology, Biarritz, p.118-125, juin 1997.
- [RADW91] K Radwan, F Foussier, C Fluhr. Multilingual access to textual databases. Proceedings of the Conference on Intelligent Text and Image Handling RIAO 91, Elsevier, p475-489, Avril 1991.
- [RASS94] A.M. Rassinoux, R.H. Baud, J.R. Scherrer. A multilingual analyser of medical texts. Second International Conference on Conceptual Structures ICCS , Springer-Verlag, p.84-96, 1994.
- [RIJS86] C.J. van Rijsbergen. A new Theoretical Framework for Information Retrieval. Proceedings of the ACM-SIGIR conference on Research and Development in Information Retrieval, Pisa, September 1986.
- [SALT94] Gerard Salton, James Allan, Chris Buckley. Automatic structuring and retrieval of large text files. Communications of the ACM, Vol.37, N°2, p.97-108, Février 1994.
- [SOWA84] J.F.Sowa. Conceptual Structures : information processing in mind and machine. The System Programming Series, Addison Wesley publishing Company, 1984.