

## Les relations sémantiques : du linguistique au formel

Nathalie Aussenac-Gilles & Patrick Séguéla \*

*La modélisation des relations sémantiques à partir de leur observation en corpus requiert une collaboration interdisciplinaire associant linguistique et ingénierie des connaissances. L'étude des manifestations linguistiques des relations dans les énoncés débouche sur la définition de marqueurs. La mise au point de patrons informatiques permet d'automatiser partiellement la recherche de traces de relations sur un corpus. Notre expérience de collaboration interdisciplinaire a débouché sur la mise au point d'un système d'aide à l'extraction et à la modélisation de relations à partir de corpus, Caméléon, que nous présentons.*

*Modelling semantic relationships from their identification in a corpus requires a cross-disciplinary collaboration between linguists and knowledge engineers. In particular, both of the two disciplines should be involved in defining patterns that characterise the linguistic behaviours revealing a given relationship in a corpus, and in tuning operational patterns for semi automatic relation extraction from electronic corpora. We report our own experiment of cross-disciplinary collaboration and present the Caméléon system that assists relation extraction and modelling from corpora.*

---

\* IRIT, Toulouse.

## 1. Introduction

Face à la croissance des informations produites à l'écrit sur support informatique, la demande est de plus en plus forte d'applications permettant de les exploiter. Les objectifs peuvent être de rechercher des informations dans les textes, de mieux les consulter ou de construire des applications ou des modèles basés sur les connaissances qu'ils contiennent. Associer les textes sources aux modèles conceptuels en assure une meilleure lisibilité et maintenance. Les textes, supports d'échanges au sein d'une communauté technique ou scientifique, sont en effet des sources de connaissances stabilisées dont l'exploitation en partie automatisée permet de recueillir rapidement un noyau de connaissances structurées. Ils sont un point de départ pour identifier les termes utilisés dans un domaine et le sens de ces termes à travers les relations lexicales qui les associent et les interdéfinissent. Les relations jouent donc un rôle capital pour accéder aux définitions des concepts tels que les acteurs du domaine les expriment dans les textes.

C'est pourquoi nous nous focalisons sur la notion de relation sémantique et faisons des propositions pour l'identification de ces relations en corpus. Notre propos est de souligner les atouts d'une approche interdisciplinaire associant linguistique et ingénierie des connaissances dans cette perspective. En effet, il nous semble indispensable d'appuyer les démarches et outils de modélisation des informaticiens sur des travaux de linguistique tant pour caractériser et repérer des relations en corpus que pour les utiliser à des fins de modélisation conceptuelle. La référence linguistique permet, par rapport à des approches statistiques par exemple, de mieux aborder la sémantique des relations, et donc d'extraire plus rapidement une relation lexicale étiquetée entre des termes. Elle autorise également de traiter des corpus peu volumineux et moins homogènes. Dans cet article, nous montrons l'intérêt de disposer des indices fournis par la linguistique pour trouver des relations entre termes utiles dans une visée de modélisation.

Etudier le problème de la modélisation de relations conceptuelles à partir de l'observation de relations lexicales en corpus permet d'aborder toutes les facettes de la sémantique de ces relations. A travers les problèmes soulevés, nous voulons utiliser au mieux les résultats de la linguistique, de la terminologie, de l'ingénierie des connaissances ou de l'intelligence artificielle. Pour cela, nous avons repéré les étapes significatives dans le processus qui va d'un texte à un modèle formel. Pour conduire chacune d'elles, nous avons cherché à fournir des éléments de réponse, tirés de ces différentes disciplines.

La notion de marqueur, définie comme le moyen de repérer des formes linguistiques faisant partie de catégories prédéfinies (grammaticales, lexicales, syntaxiques ou sémantiques), est au cœur de l'étude. Leur observation et leur définition en corpus pour révéler des relations lexicales est l'objet de recherches en linguistique. Les travaux les plus récents mettent l'accent sur la variabilité de ces marqueurs en fonction du type de corpus, du

domaine et du type de la relation. Leur utilisation systématique ne peut se faire qu'à l'aide d'un logiciel qui automatise au moins leur projection sur un corpus. Nous présentons le système Caméléon<sup>1</sup>, qui s'appuie sur l'utilisation de marqueurs pour rechercher des relations lexicales en corpus puis guider leur intégration dans un modèle. Il facilite la mise au point et le repérage de marqueurs spécifiques sur un corpus donné. Caméléon guide aussi la projection de ces marqueurs sur le corpus pour repérer des relations lexicales entre termes. Enfin, une interface de modélisation facilite l'interprétation des relations lexicales en contexte et leur exploitation pour enrichir un modèle conceptuel.

L'article s'organise donc autour de 3 parties. La première rappelle l'intérêt de se focaliser sur les relations sémantiques à partir de leurs traces en corpus pour construire des ressources sémantiques comme les terminologies ou les modèles conceptuels. La deuxième partie fait état de résultats en linguistique de corpus relatifs à l'étude des relations conceptuelles et qui nous semblent intéressants pour la modélisation des connaissances. Enfin, dans la troisième partie, nous montrons comment ces principes linguistiques guident la spécification d'une méthode et d'un outil d'aide au repérage et à la modélisation des relations conceptuelles : le système Caméléon.

## **2. Etude des relations en terminologie et en ingénierie des connaissances**

Comme l'ingénierie des connaissances, la terminologie peut viser l'objectif de constituer des ressources décrivant plus ou moins précisément et formellement des connaissances d'un domaine. Les résultats sont représentés différemment selon les applications visées. Les formalismes de référence en la matière, les réseaux sémantiques, peuvent en effet être vus soit comme de simples modélisations conceptuelles soit comme des systèmes formels. Dans tous les cas, les corpus constituent des sources de connaissances de plus en plus utilisées, et les relations sémantiques sont au cœur de la définition des termes ou concepts. L'étude des relations en corpus devient donc une activité centrale pour la construction de ces ressources.

### **2.1. Des relations pour donner du sens en terminologie**

La terminologie et les sciences de la documentation ont doublement évolué, tant par la nature des résultats produits que par les pratiques qu'elles mettent en œuvre. Aujourd'hui, des relations explicites entre termes sont de plus en plus souvent intégrées aux terminologies pour compléter les définitions en langage naturel. Ainsi, la structure des terminologies se trouve enrichie d'un niveau conceptuel avec des relations sémantiques entre concepts. Par ailleurs, la pratique terminologique accorde une plus grande importance aux textes,

<sup>1</sup> Caméléon a été développé par P. Séguéla à l'IRIT (UPS, Toulouse) et au CEA (CEN de Cadarache), en collaboration avec l'ERSS (UTM, Toulouse).

essentiellement sur support informatique. La définition des termes est désormais le résultat d'une analyse systématique de l'usage des termes en corpus. De fait, les terminologues sont demandeurs de nouveaux outils qui les aident à repérer des termes, puis à les associer à des concepts pour les définir.

Ainsi, l'identification de termes et de relations sémantiques en corpus devient un élément important de la pratique terminologique. En s'appuyant sur des textes, on espère construire la terminologie partagée par cette communauté, au-delà des différences individuelles. La difficulté est donc de repérer des relations jugées stables et consensuelles qui apparaissent explicitement ou non dans les textes. Des validations liées aux expertises du domaine et à l'usage prévu de la terminologie sont ensuite nécessaires pour en ajuster le contenu.

## 2.2. Relations conceptuelles et formalisation des connaissances en intelligence artificielle

En informatique, les relations conceptuelles sont utilisées pour représenter des connaissances dans des réseaux sémantiques (Brachman 77). Ces réseaux fournissent une représentation graphique d'une conceptualisation de connaissances, s'appuyant sur une structure dont les nœuds représentent des concepts et les arcs traduisent des relations entre concepts. Une relation d'étiquette R entre deux concepts d'étiquettes A et B traduit que A et B ont la propriété d'être en relation par R.

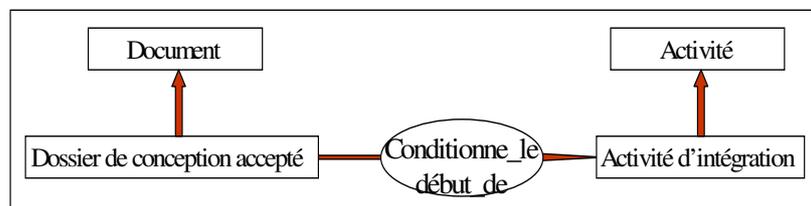


Figure 1 : Exemple de réseau sémantique et de relations conceptuelles

Par contre, les types de raisonnement possibles sur ces réseaux varient d'une implémentation informatique à l'autre (Kayser 1997). Finalement, le fait de ne pas spécifier formellement la sémantique de ces réseaux a posé des limites. Très vite, il est apparu important de différencier les concepts traduisant des classes d'objets de ceux représentant des individus, les relations prenant un sens différent selon qu'elles relient des classes ou des individus. De plus, les étiquettes de relations ne suffisant pas à les distinguer, quelques relations privilégiées, dont la relation EST-UN entre deux classes, ont été définies formellement.

En tant que mode déclaratif de représentation des connaissances, les réseaux sémantiques mettent l'accent sur la représentation des concepts ou

des actions d'un domaine, plus que sur l'explicitation de l'algorithme qui permet de raisonner à l'aide de ces connaissances. De ce fait, ils sont alternativement considérés sous deux facettes. Tantôt ils sont vus comme une notation commode pour représenter certains types de connaissances et les manipuler ; tantôt ils sont utilisés comme un langage formel, prenant un sens précis et permettant une interprétation logique par un programme. Cette dichotomie des réseaux sémantiques en souligne tout l'intérêt dans notre perspective de modélisation à partir de texte. Il est commode de les utiliser comme de simples notations graphiques à vocation d'interprétation humaine dans un premier temps, les contraintes liées à la formalisation n'étant prises en compte que dans un deuxième temps.

Ne pas formaliser d'emblée revient à choisir tardivement le modèle du monde auquel on fait référence (Kayser 1997), c'est-à-dire la bonne approximation du monde avec laquelle on pourra effectuer les raisonnements envisagés. Représenter ainsi les connaissances met en évidence la difficulté de choisir et de décrire convenablement les bons objets nécessaires pour rendre compte de connaissances et produire certains raisonnements. Lors de la construction de modèles, ce travail de choix et de description des concepts est réalisé en partie à travers l'étude des relations, depuis leur expression linguistique jusqu'à leur formalisation. Ce processus ne peut être ni systématique ni automatique puisqu'il est complètement orienté par la finalité de la modélisation.

Dans la suite de l'article, en référence aux niveaux de connaissances de Brachman (1979) et repris par Newell (1982), nous marquerons cette progression par une différence de vocabulaire. Le terme *relation conceptuelle* désignera une relation sémantique en tant que notation prenant du sens pour un individu. Par opposition, une *relation formelle* sera exprimée dans un formalisme (comme les logiques de description) qui en assure une interprétation sémantique non ambiguë pour un système formel. Passer d'une relation conceptuelle à une relation formelle revient à affiner sa sémantique. Pour cela, on associe à la relation un mécanisme d'interprétation ou alors on spécifie sa signature, c'est-à-dire le type des concepts qu'elle peut relier. Pour la relation EST-UN, l'interprétation revient souvent à gérer l'héritage des propriétés d'une classe vers ses sous-classes, puis à s'appuyer sur ces propriétés et les relations hiérarchiques pour classer de nouveaux concepts. Cette sémantique et son utilisation dans le raisonnement dépendent du langage de formalisation choisi.

### **3. Localisation de relations en corpus, en linguistique**

Nous ne prétendons pas exposer ici, à la place des linguistes, des résultats établis par leurs recherches sur l'étude des relations lexicales, mais plutôt mentionner, en tant qu'informaticiens, les principes sous-jacents à ces études

et ce qui nous semble utilisable dans la perspective de la modélisation conceptuelle.

### 3.1. La notion de marqueur et de relation lexicale

En linguistique, il est admis de considérer que la langue répond à des règles que l'on peut expliciter. La notion de marqueur traduit cette hypothèse relativement au repérage de relations sémantiques. On suppose que l'observation de certaines formes linguistiques entre deux ou plusieurs éléments du lexique peut révéler un rapport de sens entre ces éléments. Il s'agit alors d'élaborer des marqueurs pour rendre compte précisément de fonctionnement lexicaux et les associer à une interprétation sémantique systématique. Deux hypothèses sous-tendent cette approche : 1) la possibilité, à travers des éléments de forme, d'accéder à du contenu ; 2) le contenu identifié est une relation binaire entre concepts.

Un *marqueur* correspond à une *formule linguistique* dont l'interprétation définit régulièrement le même rapport de sens entre des termes. Nous appelons *relation lexicale* le type de relation entre ces deux termes. Inversement, le marqueur caractérise la relation à laquelle il est associé et, en poussant à l'extrême cette idée, on peut dire que l'ensemble des marqueurs d'un type de relation est une manière très explicite de lui donner du sens, plus que sa seule étiquette. En effet, les marqueurs mettent en évidence tout ce qui est commun à toutes les occurrences de la relation.

Les marqueurs sont composés de mots de la langue et de symboles renvoyant à des catégories grammaticales (SN pour Syntagme Nominal, Nom, Dét. pour Déterminant, Adj. pour Adjectif, etc.), à des classes sémantiques (i.e. *NomOutil* pour retrouver *logiciel, atelier, outil, plateforme, ...*) ou, plus rarement, à des rôles argumentaux (agent, patient, objet, ...). Par exemple, la formule linguistique suivante est un des marqueurs de la relation d'hyponymie (SN2 EST-UN SN1) :

*SN1 ET ADVERBE\_DE SPECIFICATION SN2*

où SN1 et SN2 sont des syntagmes nominaux, ET renvoie au mot "et", et ADVERBE\_DE SPECIFICATION à une des expressions adverbiales suivantes : "plus précisément", "tout particulièrement", "plus particulièrement", "en particulier", "surtout", "principalement", "essentiellement", "notamment"

Il permet de retrouver entre autres la phrase suivante :

*On calcule la thermique de la partie haute du réacteur, et en particulier la température du toit.*

Cette phrase révèle une relation d'hyponymie entre les syntagmes nominaux *thermique de la partie haute du réacteur* et *température du toit*.

L'étude des relations par les linguistes entre donc dans une perspective sémantique. Il s'agit de rendre compte de régularités de la langue, de proposer des méthodes pour trouver de nouveaux marqueurs pour une

nouvelle relation et de caractériser les corpus dans lesquels on les trouve. La pertinence des marqueurs est validée en fonction de leur efficacité pour retrouver des contextes pouvant contenir des relations d'un certain type.

### **3.2. Résultats disponibles**

#### **3.2.1. Marqueurs généraux et relations générales**

Pratiquement, les résultats de linguistique sur les marqueurs associés à des relations lexicales sont bien avancés pour ce qui concerne les relations de méronymie et d'hyperonymie. Ainsi, les recherches sur l'hyperonymie en français (A. Borillo (1996) en linguistique ou C. Jouis (1993) dans SEEK en linguistique informatique), constituent des ressources importantes de marqueurs de la relation *est\_un*. De même, les articles de G. Otman (1996) ou A. Jackiewicz (1996) sur la méronymie fournissent des marqueurs lexicaux de la relation *partie-tout* en français. Ces marqueurs permettent de définir des patrons de fouille de corpus. Deux écoles ont cours à ce jour pour juger de leur validité :

- L'une fait de la régularité une hypothèse très forte, au point de supposer que les marqueurs d'un type de relation sont universels et pourraient être observés sur tout corpus ; ces travaux étudient dans des textes très variés des relations elles-mêmes indépendantes des corpus : l'hyperonymie ou la méronymie.
- L'autre nuance cette idée et s'attend à trouver des spécificités liées à la cohérence du corpus (domaine très technique, sujet précis, etc.). Cette cohérence justifie l'attente de relations propres à un corpus, voire la présence de comportements linguistiques déviants (polysémies fixées sur un domaine, nouveaux sens associés à certaines formes) par rapport aux connaissances a priori sur la langue.

Nous nous référons à la deuxième école, puisque nous travaillons sur des corpus dont nous cherchons à extraire un maximum de connaissances. L'idée est d'exploiter au maximum des marqueurs génériques, mais aussi de les adapter et de repérer de nouveaux marqueurs pour des relations spécifiques. C'est alors la démarche de repérage et de mise au point de marqueurs qui peut être considérée comme un résultat. Cependant, l'informaticien se trouve parfois démuné pour mettre en œuvre ce type d'approche pour deux raisons essentielles :

- les marqueurs disponibles à ce jour sont éparpillés dans la littérature, rarement répertoriés de manière systématique y compris par les auteurs qui les ont trouvés ou utilisés ;
- les contextes de validité des marqueurs ne sont pas toujours suffisamment précisés, les linguistes considérant comme un travail de linguistique à part en entière l'évaluation de marqueurs connus ou la mise au point d'un nouveau marqueur sur un corpus particulier.

Finalement, le besoin se fait sentir d'une gestion plus systématique de ces résultats, par leur capitalisation dans une base de données par exemple.

### 3.2.2. Variabilité du lien marqueur/relation : influence du corpus

Un des résultats de notre collaboration avec des linguistes de l'Équipe de Recherche en Syntaxe et Sémantique (ERSS) de Toulouse est d'avoir mis en évidence (Condamines 1999, Rebeyrolle 2000) qu'une hypothèse sous-tend l'observation et la définition de marqueurs en corpus et leur capacité à révéler des relations lexicales. Il existe en effet, dans le triplet corpus / relation / marqueur, des liens de dépendance étroits qu'il faut prendre en compte. La maîtrise de ces liens pourrait permettre de disposer de repères afin de choisir des textes pertinents pour y chercher des types de connaissances (descriptives, opératoires, etc.) en fonction des types de relations présentes. Les exemples suivants illustrent ces différentes dépendances :

- Une même relation peut prendre de multiples formes et donc s'exprimer par différents marqueurs : par exemple, d'autres marqueurs de la relation d'hyponymie sont *tous les N2 sauf det N1* et *Det N1 comme det N2* permettent de trouver la relation *Coatis EST-UN outil* dans la phrase :  
*les connaissances causales jouent donc un rôle privilégié et leur identification à l'aide d'un outil comme Coatis se révèle très utile.*
- La présence d'un type de relation dans un corpus dépend du domaine étudié : ainsi, dans un corpus de génie logiciel, on relève la relation *conditionne le début de* entre deux tâches, dont un des marqueurs (*phase, étape, processus, activité, NOMINALISATION*) + (*CONJ. DE SUBORDINATION TEMPORELLE*) permet de retrouver la phrase suivante :  
*Les activités liées à l'intégration débutent dès que le dossier de conception produit et le plan d'intégration produit sont acceptés.*
- La validité d'un marqueur dépend du corpus : ainsi, des marqueurs génériques doivent être contraints pour être plus pertinents sur certains corpus alors que d'autres en sont absents. Des marqueurs spécifiques ne se trouvent que dans certains types de corpus, comme le marqueur *chez* qui indique une relation *partie-tout* seulement dans les textes didactiques en sciences naturelles (*Chez les primates, la mandibule a des mouvements verticaux*).

### 3.3. Des marqueurs aux relations lexicales : méthodes et techniques

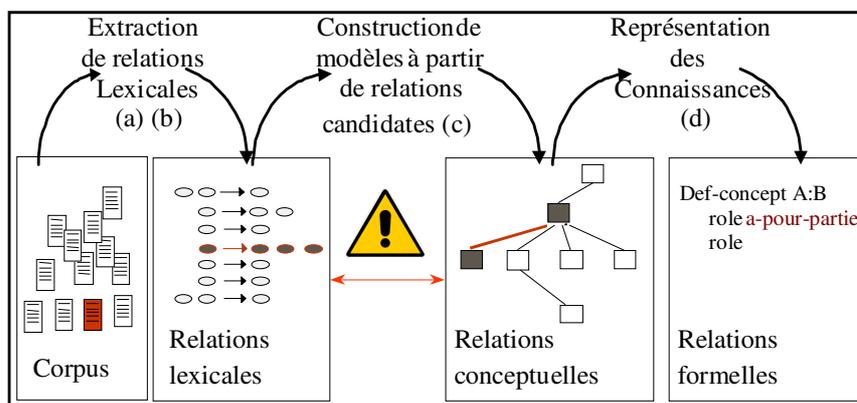
Bien sûr, il est illusoire de prétendre trouver toutes les connaissances utiles à une application, et uniquement celles-là, dans les textes. Ceci a une double conséquence : d'une part, le processus de modélisation des relations ne peut être totalement automatisé (à chaque étape, il faut décider de la pertinence ou non d'une relation) ; d'autre part, d'autres sources de connaissances, les spécifications du système à concevoir mais aussi des principes de

modélisation (comme la normalisation) doivent être pris en compte pour décider de la définition et de la structuration de concepts à l'aide des relations trouvées.

### 3.3.1. Du corpus au modèle : un processus complexe

Envisager la modélisation de relations conceptuelles à partir de l'observation de relations lexicales en corpus permet d'aborder toutes les facettes d'étude de ces relations et, à travers les problèmes soulevés, d'utiliser au mieux les résultats des différentes disciplines qui s'y intéressent. En effet, on peut repérer quatre étapes significatives dans le processus qui va d'un texte à un modèle formel (figure 2) :

- (a) caractériser les comportements linguistiques stables associés à des relations lexicales, et utiliser ensuite cette caractérisation pour repérer des relations spécifiques en corpus ;
- (b) à partir de l'expression lexicale brute d'une relation spécifique, décider d'une relation ayant une étiquette donnée entre deux termes ; le fait de la nommer revient à la classer dans une typologie de relations ;
- (c) passer ensuite du niveau linguistique au niveau conceptuel, en décidant du devenir de cette relation au sein d'un réseau de concepts associés par des relations étiquetées ; cette relation contribue ainsi à la définition des concepts et à leur différenciation par rapport à d'autres ;
- (d) représenter ces relations conceptuelles selon une représentation qui en garantisse une interprétation unique et non ambiguë par un système formel.



*Figure 2 : Les relations sémantiques : du corpus au modèle conceptuel*

L'étape (a) comporte en fait deux facettes : un travail purement linguistique conduit à définir les marqueurs d'un point de vue linguistique, à l'aide de primitives sémantiques, syntaxiques et lexicales ; les utiliser pour repérer des relations en corpus suppose l'opérationnalisation de ces marqueurs et l'utilisation d'un logiciel adapté. Or trouver la bonne forme informatique d'un marqueur est loin d'être trivial. Sa qualité conditionne la capacité du marqueur à retrouver un type de relation dans un corpus. De même, la manière de conduire les étapes suivantes ne peut émerger que d'une confrontation interdisciplinaire.

### 3.3.2. Une méthode linguistique

Une méthode a été proposée et expérimentée par A. Condamines et J. Rebeyrolle (2000) pour construire un réseau conceptuel à partir de textes selon des principes linguistiques. Elle consiste à exploiter successivement les résultats de deux outils d'analyse de textes, un extracteur de termes, Lexter, et un concordancier pour le repérage de relations, Sato. Nous nous intéressons bien sûr aux propositions faites quant au repérage des relations, étape jugée cruciale pour déterminer si un terme sera ou non retenu dans le modèle.

Ces auteurs insistent sur la nécessité de différencier l'étude des relations génériques qui peuvent *a priori* se retrouver dans tout corpus de l'étude des relations spécifiques à ce corpus. Il s'agit de commencer par utiliser des connaissances générales (des marqueurs) sur les relations, de les projeter sur le corpus pour les évaluer, puis de les adapter si besoin et de les tester à nouveau sur le corpus. Ce travail très dynamique permet d'aboutir à la bonne formulation de patrons de fouille pour des relations générales qui soient efficaces sur ce corpus, c'est-à-dire ramenant peu de mauvais exemples (bruit réduit) tout en étant le plus exhaustif possible (silence réduit).

Au passage, la lecture des occurrences permet aussi de repérer des relations syntagmatiques spécifiques au corpus et d'identifier les marqueurs associés, qui sont mis au point selon le même principe itératif de tests par projection. Ce processus de généralisation suppose des compétences linguistiques pour parvenir à une bonne caractérisation des marqueurs à partir de la lecture des occurrences.

### 3.3.3. Opérationnalisation des marqueurs

L'exploitation de marqueurs pour la recherche de relation en corpus revient à s'appuyer fortement sur l'hypothèse de régularité pour retrouver des connaissances, plus précisément des relations conceptuelles et les concepts reliés. La terminologie et l'informatique (traitement du langage naturel et ingénierie des connaissances) détournent ainsi les marqueurs de leur rôle

initial pour en faire des outils de recherche de connaissances spécifiques et non pour rendre compte de l'expression de ces connaissances.

Dans la suite, nous marquerons la différence entre la *formule linguistique* et le *schéma informatique* qui correspond à son opérationnalisation. Comme le soulignent A. Condamines et J. Rebeyrolle (2000) ou P. Séguéla (1999), l'opérationnalisation directe de marqueurs donne des patrons peu précis (ils ramènent des exemples erronés et en oublie une grande partie) et des résultats peu satisfaisants en corpus. Outre l'ajustement du marqueur au corpus, il faut trouver le schéma le plus efficace de ce marqueur spécifique. En effet, un même marqueur sera opérationnalisé différemment en fonction des capacités du logiciel utilisé et du corpus. Le type d'outil de traitement automatique des langues et la richesse du langage informatique dont on se dote ont un impact significatif sur le type de recherche possible sur le corpus. Par exemple, l'analyse syntaxique préalable du corpus peut permettre de définir des marqueurs utilisant des indications syntaxiques.

Pour un langage et un outil donnés, la mise au point d'un marqueur sur un corpus consiste donc à fixer le schéma qui obtient les meilleurs résultats après évaluation et à décider du type de relation qu'il révèle. Projeter un marqueur consiste ensuite à lancer un programme qui recherche toutes les occurrences du schéma dans le corpus, les enregistre ou les présente à l'utilisateur. La définition de patrons de fouille vraiment efficaces requiert des choix, méthodologiques et techniques, et représente un travail complexe associant informatique et linguistique, comme celui décrit par J. Rebeyrolle et L. Tanguy (2001) dans cet ouvrage.

### **3.3.4. Des logiciels pour la recherche de relations**

Plusieurs types d'outils informatiques permettent d'utiliser la notion de marqueur pour extraire des relations de corpus sur des textes en français. Tous s'appuient sur un processus semi-automatique : ils retournent des contextes contenant potentiellement une relation, que nous appelons *hypothèses de relations* dans la suite. Chaque contexte doit être lu et validé pour plusieurs raisons : un marqueur n'est jamais complètement fiable, il se peut que le contexte retourné ne contienne pas la relation cherchée ; ensuite, suivant l'application visée et le modèle à construire, la relation signalée n'est soit pas valide, soit pas pertinente.

Parmi ces outils, on trouve les concordanciers, qui cherchent toutes les occurrences en corpus d'une série de mots fixés par l'utilisateur. Certains de ces systèmes, comme SATO (Daoust), sont plus performants parce qu'ils utilisent la caractérisation de mots par leurs propriétés syntaxiques ou leur famille sémantique. En effet, ces systèmes s'appuient sur des corpus analysés

et étiquetés syntaxiquement, ainsi que sur des dictionnaires sémantiques regroupant certains mots du domaine par famille sémantique. Basés sur le même principe, Yakwa (Tanguy 2000) et plus récemment Linguae (Szulman 2001) sont spécialisés pour définir des schémas à partir de formules linguistiques marquant des relations. Ils permettent en particulier de gérer les marqueurs utilisés et leurs résultats.

D'autre part, il existe des outils dédiés à la recherche de certains types de relations à partir de marqueurs : SEEK (Jouis 1993) pour les relations EST-UN et PARTIE-DE, Prométhée (Morin 1999) pour la relation EST-UN, Coatis (Garcia 1998) pour les relations de causalité entre événements. En fait, ces logiciels s'appuient sur une des deux approches suivantes, tout à fait complémentaires :

- partir de marqueurs pour trouver des relations lexicales et, ainsi, des couples de termes en relation ; des systèmes comme SEEK ou Coatis disposent d'une base de marqueurs génériques reconnus pertinents sur différents types de corpus, mais non modifiables ;
- partir de couples de termes dont on sait qu'ils sont en relation (ou plus précisément de la traduction linguistique de concepts), ou de l'étude de termes fortement co-occurents (de segments répétés), pour étudier la régularité de leurs contextes et en dégager des marqueurs spécifiques au corpus étudié. C'est le mode de fonctionnement de Prométhée, qui recherche des couples à partir de l'étude statistique de cooccurrences de termes, et de LIKEs (Rousselot *et al.* 1996), qui recherche des segments répétés.

L'utilisation successive de ces deux approches définit un cycle de repérage de marqueurs et de relations propres au corpus. Il sert de base en partie à la méthode et au logiciel Caméléon présenté dans la suite. L'originalité de Caméléon est de couvrir la totalité du processus de modélisation, depuis la mise au point des marqueurs jusqu'à l'enrichissement du modèle conceptuel à l'aide de concepts et de relations extraits des textes.

#### 4. Un logiciel d'aide à l'extraction et à la modélisation de relations conceptuelles à partir de textes

##### 4.1. Présentation

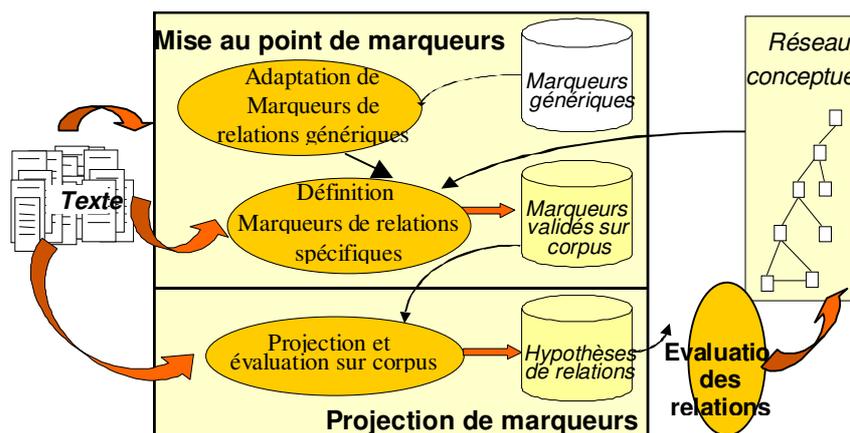


Figure 3 : Principales fonctionnalités de Caméléon

Le système Caméléon aide au repérage de relations dans des textes techniques et à leur intégration dans un modèle. Il a été développé dans le contexte du système REX de gestion de retour d'expérience, dans lequel un ensemble de modèles du domaine organise termes et concepts pour faciliter la recherche d'information dans des textes rédigés pour enregistrer la mémoire d'un projet. Caméléon comprend une méthode et un logiciel support qui assistent trois types d'opérations (Figure 3) :

- la mise au point d'une base de marqueurs et de relations adaptés à un corpus à partir d'une base de marqueurs génériques et de leur étude en corpus ;
- leur projection sur le corpus pour localiser des hypothèses de relations lexicales entre termes.
- la validation de ces hypothèses en contexte et leur sélection pour enrichir un modèle conceptuel.

Les marqueurs génériques sont tirés de travaux de linguistique sur les relations d'hyponymie et de méronymie, alors que les marqueurs spécifiques sont définis à l'aide de patrons trouvés par l'utilisateur à partir de l'étude des contextes de cooccurrence de termes en relation. En effet, la démarche prévoit de travailler sur des textes techniques, dans lesquels on s'attend à trouver des phénomènes linguistiques typiques : (a) des interactions spécifiques entre les objets du domaine qui se traduisent par des types de relations propres à ce corpus ; (b) des usages spécifiques en corpus,

qui fixent le sens de termes polysémiques, et aussi qui permettent de trouver des marqueurs spécifiques de relations génériques. Ces phénomènes justifient la phase de mise au point de marqueurs.

Pour faciliter la modélisation, Caméléon permet ensuite de reprendre les hypothèses de relations lexicales en contexte grâce à une interface d'enrichissement de modèle conceptuel. Dans sa version actuelle, ce module permet de compléter des modèles Rex. Ainsi, le passage de la relation lexicale à une relation conceptuelle prend en compte la place du concept concerné par une relation dans le modèle existant, et les relations qu'il entretient déjà avec d'autres concepts. Une relation conceptuelle n'est ajoutée que si la relation lexicale permet de mieux définir le concept de manière cohérente avec les concepts qui lui sont directement associés et avec l'objectif pour lequel le modèle est conçu. La personne chargée de la modélisation donne ainsi à l'indice linguistique une pertinence cognitive.

#### 4.2. Motivations

La définition de Caméléon est essentiellement motivée par la nécessité de disposer d'un environnement qui couvre tout le processus de modélisation, depuis l'analyse de textes techniques à l'aide de logiciels d'extraction jusqu'à l'intégration dans un modèle des connaissances tirées de ces textes. En effet, à ce jour, les travaux de traitement automatique des langues et d'ingénierie des connaissances étaient fortement découplés. D'un côté, on trouve les systèmes d'extraction de termes ou d'aide à l'extraction de relation mentionnés en partie 3. D'un autre côté, on dispose de plates-formes ou de langages de modélisation de réseaux conceptuels. Or le fait de viser une application finalisée oblige à mieux intégrer ces deux composantes, à anticiper le fait que les résultats servent dans le cadre d'une modélisation dès l'analyse de textes. C'est là l'originalité essentielle des propositions méthodologiques et du logiciel Caméléon.

188

Caméléon répond donc à des besoins de deux ordres, fondamentaux et pratiques :

- 1) Fournir des hypothèses de relations riches sémantiquement, c'est-à-dire en faisant des propositions sur l'étiquette de la relation ainsi que sur les termes arguments. Cette volonté a conduit d'une part à choisir une approche basée sur l'utilisation de marqueurs au détriment d'autres méthodes comme l'étude des distributions de contextes autour des termes, et d'autre part à utiliser un extracteur de terme au préalable de la recherche de relations. De plus, la forme de ces hypothèses doit être commode pour faciliter la modélisation.
- 2) Il a été décidé de thésauriser les résultats en linguistique sur les relations d'hyponymie et de méronymie dans une base de marqueurs génériques, applicable sur tout corpus. Cette base doit assurer une

- gestion pérenne de marqueurs proposés par la linguistique, les synthétiser, les rendre accessibles et facilement modifiables.
- 3) Les travaux sur les relations dans les textes techniques ont justifié de pouvoir spécialiser les marqueurs et les types des relations extraites du corpus : la définition de nouveaux marqueurs et de nouvelles relations doit donc être simple et même favorisée. Pour cela, on soumet au cognicien des contextes de cooccurrences de termes dont on sait qu'ils sont déjà en relation. Ces contextes peuvent révéler des marqueurs ou des relations spécifiques.
  - 4) Proposer un langage simple pour l'expression des marqueurs : pour cela, un langage d'expressions régulières a été défini. Cette volonté a aussi conduit à ne pas s'appuyer sur une analyse syntaxique préalable du corpus. D'autres raisons motivent aussi ce choix : au moment de spécifier Caméléon, aucun analyseur n'était facilement disponible, de qualité suffisante pour traiter des corpus non accentués et très techniques, comme ceux du CEA.
  - 5) Tracer le processus de validation des marqueurs dans des fichiers : il semble nécessaire de conserver les différentes formes testées pour la mise au point d'un marqueur, l'évolution de l'ensemble des marqueurs et de l'ensemble des relations pour un corpus donné.
  - 6) Présenter les résultats (hypothèses de relations) dans le contexte du modèle afin de décider de les y insérer. Le choix des termes / la définition de nouveaux concepts dans un modèle ne se fait pas par seule lecture des termes, mais aussi en fonction des relations déjà existantes dans le modèle et en fonction de l'application cible.

#### **4.3. Mise au point d'une base de marqueurs avec Caméléon**

La base de marqueurs de relations génériques de Caméléon ainsi que les marqueurs et relations spécifiques sont gérés dans des fichiers structurés. Caméléon permet d'adapter l'association entre marqueurs et relations à chaque corpus étudié. Les relations sont représentées par leur seule étiquette et organisées au sein d'une hiérarchie. Cette hiérarchie offre un cadre pour ajouter de nouvelles relations ou pour associer relations et marqueurs. Les marqueurs sont désignés également par une étiquette et caractérisés par un schéma informatique, une expression régulière simple qui sera interprétée lors de la projection sur le corpus. Pour un corpus donné, la mise au point d'une base de marqueurs comprend deux étapes : l'évaluation et l'adaptation de marqueurs génériques ; le repérage et l'évaluation de nouveaux marqueurs pour des relations génériques ou spécifiques.

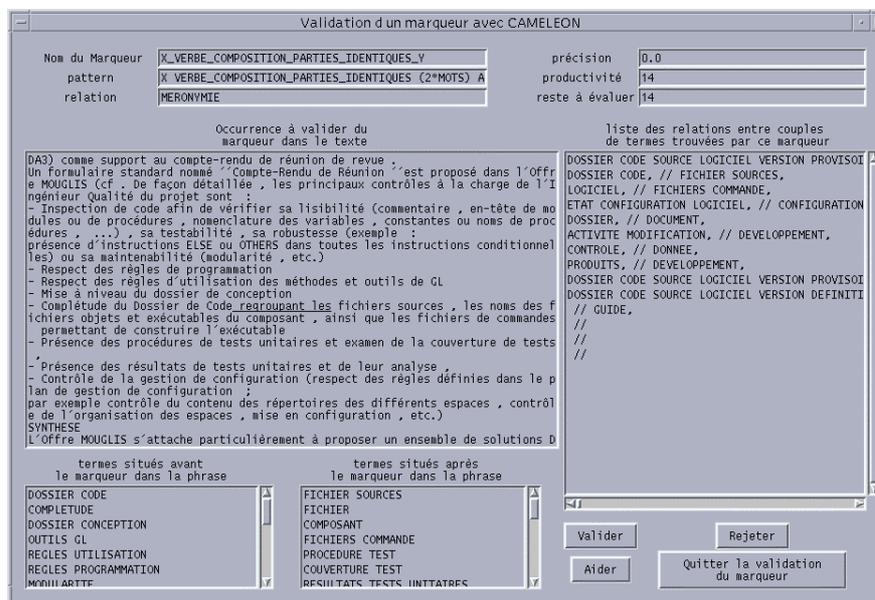


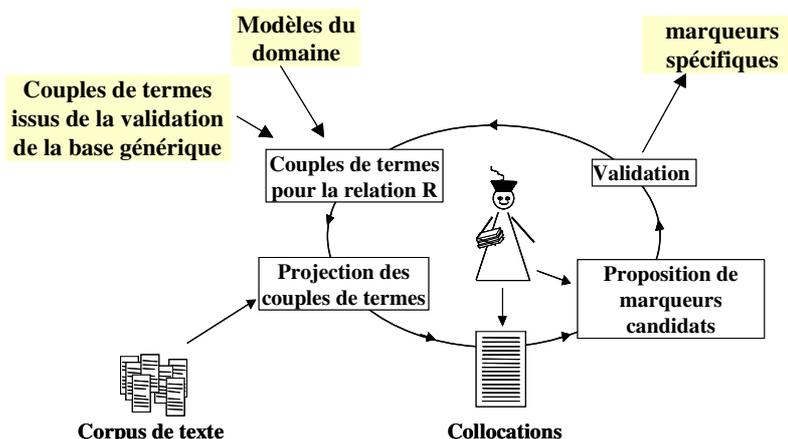
Figure 4 : Interface de validation de marqueur dans Caméléon

#### 4.3.1. Evaluation des marqueurs de la base générique

Cette étape consiste à juger, pour chaque marqueur de la base générique, si la relation associée est présente ou non dans un échantillon de ses occurrences pour décider de conserver, rejeter ou modifier ce marqueur. L'interface de la figure 4 est le support de cette validation : la liste des couples trouvés par le marqueur est proposée à droite et, pour chacun d'un, on peut lire l'occurrence contenant le marqueur (cadre de gauche) puis décider de valider ou non cette hypothèse de relation (boutons en bas à droite). Le nombre d'hypothèses de relations validées sur le nombre d'occurrences consultées parmi celles trouvées par le marqueur permet d'associer une précision chiffrée au marqueur. Sa valeur est un repère pour guider le cognicien qui doit décider de retenir ou non ce marqueur.

La base générique comporte des marqueurs de relations d'hyponymie (33 marqueurs), de méronymie et ses sous-types (122 marqueurs répartis en 7 types de relations). Il est important d'avoir en tête que certains de ces marqueurs sont très peu occurrents en général, que d'autres sont plus ou moins productifs selon le type de corpus. Enfin, certains marqueurs méritent d'être spécialisés pour être plus performants sur un corpus donné. Lors des expériences que nous avons menées, les bases validées comprenaient près de la moitié des marqueurs initiaux.

#### 4.3.2. Projection de couples de termes reliés pour le repérage et la mise au point de marqueurs spécifiques



*Figure 5 : Processus cyclique d'identification de marqueurs spécifiques*

L'état courant du modèle ainsi que l'étape de validation des marqueurs génériques fournissent des couples de termes en relation : ce sont les concepts déjà reliés dans le modèle, et les termes présents dans les occurrences validées par le cogniticien. Ces couples et les relations qui les relient sont enregistrés dans un fichier qui sert d'entrée à Caméléon pour rechercher leurs contextes de cooccurrence (phrases dans lesquelles deux termes reliés sont présents). Le cogniticien peut ensuite, pour chaque type de relation ou chaque couple de terme, scruter les contextes à la recherche de nouveaux marqueurs, ou d'une meilleure formulation de marqueurs génériques, ou de marqueurs de nouveaux types de relations. Il enrichit alors la base de marqueurs validés et projette ces nouveaux marqueurs sur le corpus. Il les évalue et peut trouver ainsi de nouveaux couples de termes. Un processus cyclique se met en place, qui se stabilise assez rapidement dès que l'on ne trouve plus de nouveau couple en relation.

## 5. Des relations à un modèle conceptuel

### 5.1. Différentes exigences pour différentes applications

L'exploration de corpus pour y rechercher des connaissances peut se faire selon de multiples perspectives, qui se différencient par des paramètres aussi variés que la taille des corpus étudiés, le fait de savoir caractériser ce que l'on cherche ou non, le besoin de rendre compte fidèlement du texte ou non, la nature du modèle à construire à partir du texte, etc. Pour chacune d'elles, les exigences sur les qualités des marqueurs sont spécifiques :

- recherche d'information : ces techniques tolèrent des marqueurs peu précis sur les relations, mais précisant les classes sémantiques recherchées ;
- apprentissage automatique à partir de texte : dans ce cas, on souhaite disposer de marqueurs très fiables, robustes, déjà validés ;
- acquisition de connaissances à partir de textes : l'intervention d'un cogniticien pour valider et filtrer les résultats des marqueurs en fonction de l'application cible autorise une moindre fiabilité de ceux-ci ;
- modélisation du contenu de textes, pour construire une base de connaissances terminologiques par exemple : dans cet objectif, l'analyse linguistique, très poussée, est garante du contenu du modèle final ; les marqueurs sont systématiquement remis en question par observation de toutes leurs occurrences en corpus.

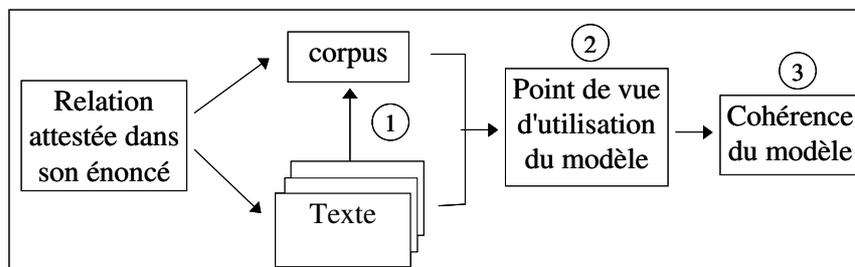
Par contre, dans tous les cas, l'exploitation des occurrences des marqueurs requiert un filtrage. Ce filtrage doit pouvoir tenir compte des termes présents dans les occurrences, de leur qualité (donc de la précision du marqueur qui les a trouvées) ou encore d'un type de relation. Il permet de valider les hypothèses de relations selon différents points de vue qui varient au fur et à mesure de cette validation. On peut travailler par exemple sur toutes les hypothèses trouvées pour un terme donné, de la plus précise à la moins précise, pour une relation donnée.

Le paragraphe suivant décrit comment les relations lexicales peuvent servir à alimenter un modèle conceptuel, et le problème que soulève leur validation. Nous illustrerons ensuite l'utilisation de Caméléon pour valider les propositions de relations trouvées par marqueur puis les intégrer dans un modèle conceptuel.

## 5.2. Des relations lexicales à un modèle conceptuel : problèmes soulevés

### 5.2.1. Niveaux d'interprétation des relations

La décision d'intégrer une relation conceptuelle dans un modèle à partir de l'observation en corpus d'une relation lexicale ne peut être prise directement. La complexité de la décision est due autant à l'hétérogénéité du corpus qu'à la distance qui sépare le corpus de l'application visée. Pour bien comprendre la nature de cette décision, décomposons-la en fonction des contextes, parfois contradictoires, qui peuvent être pris en compte. En fait, il s'agit de filtres qui viennent réduire les chances de conserver une relation observée. Nous différencions (fig. 6) l'*énoncé*, c'est-à-dire la phrase ou le paragraphe englobant l'occurrence, le *texte* auquel elle appartient, et le *corpus* qui regroupe différents textes. Nous supposons le texte homogène et cohérent, ce qui le démarque du corpus, qui, lui, l'est moins. L'*application* vient se rajouter comme un quatrième filtre, pragmatique, combiné à la contrainte de conserver la cohérence du modèle.



**Figure 6** : Niveaux d'interprétation d'une hypothèse de relation

De fait, l'interprétation d'une relation observée comprend une série d'analyses :

- une validation de la relation dans l'énoncé dont elle est issue : le marqueur doit bien renvoyer à la relation, et les termes désigner des concepts étudiés ;
- un passage de l'énoncé au texte : on cherche à généraliser au texte la relation validée localement, selon les conditions de vérité des autres connaissances déjà établies à partir du même texte ;
- une validation en corpus : relativement direct si le corpus est homogène (les textes sont cohérents et donc les relations restent valides), le passage du texte au corpus peut être complexe ((1) sur la figure 6) ;
- du corpus au modèle : il s'agit de tenir compte de l'état courant du modèle et des objectifs d'utilisation du modèle ((2) Fig. 6) pour décider d'y ajouter une relation trouvée, tout en respectant la cohérence du modèle (3) et les critères de modélisation retenus jusque-là. Ce processus revient à une normalisation.

Terminologies et modèles conceptuels se démarquent d'ailleurs par le degré de cohérence des connaissances qu'ils représentent : le terminologue pourra s'arrêter à une validation locale du terme par rapport à des termes voisins, alors qu'en ingénierie des connaissances, les concepts introduits dans un modèle doivent maintenir la cohérence globale de ce modèle.

### **5.2.2. Un processus cyclique de construction de modèle**

Caméléon propose de construire le modèle selon un processus dirigé par la personne qui modélise en fonction des objectifs de l'application, et facilité par les indices que sont les hypothèses de relations extraites automatiquement.

- Des relations lexicales pour amorcer la construction du modèle : il s'agit ici de s'appuyer sur des relations générales le plus possible et sur les concepts centraux du domaine. Ces concepts sont choisis par la

personne qui modélise, avec l'aide d'un expert ou en fonction de critères statistiques d'occurrence des termes désignant ces concepts. Comme le suggèrent A. Condamines et J. Rebeyrolle (1998), les relations paradigmatiques, comme EST-UN, fournissent un point de départ solide à la modélisation. Elles permettent de construire une ossature du modèle. La fiabilité des marqueurs associés à cette relation facilite le processus de validation, même avec peu de connaissances sur le domaine étudié. Les marqueurs de méronymie fournissent également un point de départ stable, mais déjà plus difficile à intégrer.

- Des relations lexicales pour définir des termes et décrire des concepts : partant d'un concept présent dans le modèle, on étudie les occurrences des termes qui le désignent dans le corpus lorsque ces occurrences contiennent des marques de relations lexicales. L'interprétation des occurrences permet de décider d'ajouter ou non de nouvelles relations conceptuelles dans le modèle, et éventuellement de nouveaux concepts (ceux reliés par ces relations au concept étudié). Ces concepts sont définis au cycle suivant, toujours grâce aux relations lexicales trouvées en corpus.
- Pour les linguistes, ce processus cyclique comporte également le repérage de nouvelles relations et marques linguistiques de relations en corpus. Dans le contexte de la modélisation, toutes les relations spécifiques doivent être identifiées dans la phase de mise au point des marqueurs, qui incite justement à rechercher de nouveaux marqueurs.

### **5.3. Enrichissement de modèle avec Caméléon**

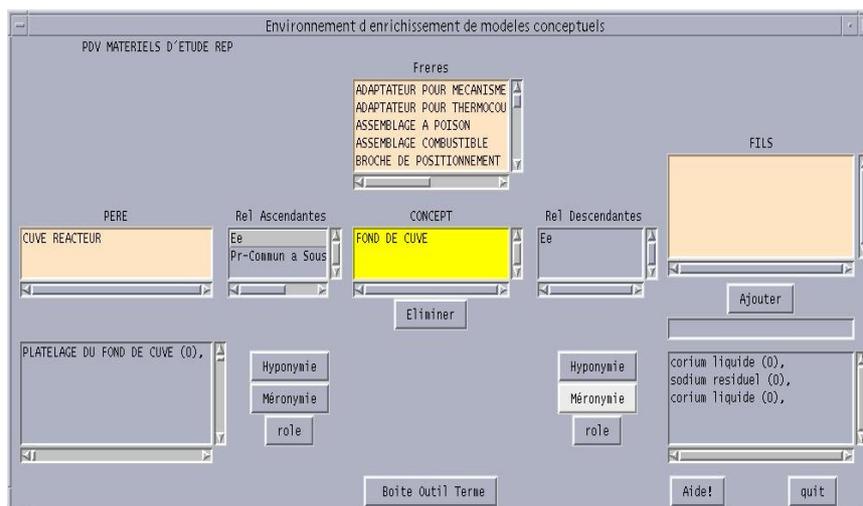
Pour un corpus et un objectif de modélisation donnés, Caméléon suggère donc de mettre au point une base de marqueurs et de types de relations pertinents pour ce corpus, de les projeter puis d'en valider une à une les occurrences pour en décider de l'intérêt avant de les intégrer ou non dans le modèle existant.

#### **5.3.1. Les hypothèses de relation**

Dans Caméléon, la projection des marqueurs correspond à une commande qui lance l'exploration du corpus à l'aide de la base validée sur le corpus étudié. Chaque identification d'un marqueur donne lieu à une hypothèse de relation, pour laquelle on enregistre le marqueur qui a permis de la trouver, le texte trouvé et le nom de la relation. Le fichier des résultats est organisé par relation et par marqueur. Il peut être exploité directement ou via l'interface prévue dans Caméléon. Dans tous les cas, ces hypothèses doivent être validées une à une pour savoir si elles donnent lieu ou non à l'ajout de connaissances dans le modèle.

### 5.3.2. Validation des hypothèses de relation pour enrichir un modèle

La validation des relations suit le principe de jugement de validité et de pertinence de l'hypothèse développé au 5.2.2. Pour le filtre final du modèle, Caméléon oriente le cogniticien de manière à l'inciter à appliquer le principe de différenciation homogène du concept : il doit posséder un trait commun avec son père, un autre avec ses frères, et au moins un trait le différenciant de son père et d'autres de ses frères. Pour cela, le concept est présenté avec ses frères et son père selon le type de relation étudié dans l'interface principale du processus d'enrichissement (figure 7). Le cogniticien choisit de compléter le modèle autour d'un concept, mentionné au centre de l'écran, et pour un type de relation particulier. La fenêtre affiche alors son concept père (à gauche), ses concepts frères (au-dessus) et ses concepts fils (à droite) pour cette relation (relation ascendante et descendante) présents dans l'état courant du modèle. La fenêtre présente aussi d'autres concepts candidats à être des fils de ce concept pour cette relation : ce sont les hypothèses trouvées pour ce terme et cette relation par Caméléon (liste en bas à droite sur la figure).



*Figure 7 : Caméléon : validation d'une relation dans le contexte du modèle*

Le cogniticien doit ensuite consulter une à une ces propositions pour les valider ou les rejeter. Pour cela, une autre fenêtre permet de lire le contexte de la relation, et confirmer les termes et les relations effectivement présents ou non dans ce contexte. Le cogniticien peut alors définir de nouveaux concepts à partir des termes reliés, et de nouvelles relations. Il poursuit ensuite la structuration du modèle à partir d'un autre concept, ou par l'étude d'un autre type de relation. Il faut noter que le modèle peut aussi être enrichi de concepts et de relations non trouvés par Caméléon.

## 6. Conclusion

Nous avons montré les enjeux relatifs au repérage puis à la modélisation de relations, ainsi que l'intérêt d'aborder ce problème par le biais de marqueurs linguistiques de relations, étudiés dans un cadre interdisciplinaire. Une collaboration interdisciplinaire de ce type, s'appuyant sur un véritable échange d'études et de savoir-faire pour contribuer à une problématique commune, nous a permis de mettre au point une méthode et un outil, Caméléon. Ce système est adapté aussi bien à une étude linguistique des marqueurs (grâce au cycle de spécialisation des marqueurs et des relations pour un corpus donné) qu'à un travail de modélisation de connaissances à partir de textes (grâce à l'interface de validation des relations en contexte). Nous envisageons de poursuivre ce travail en vue de mieux intégrer nos résultats. Par exemple, l'application pour laquelle le modèle est construit sera prise en compte plus tôt pour sélectionner des marqueurs ou des relations plus pertinents à étudier. Une autre perspective serait aussi d'affiner la sémantique des relations, encore embryonnaire.

Par ailleurs, l'outil Caméléon étant disponible, nous pouvons pousser plus loin plusieurs pistes de recherche :

- valider / évaluer des marqueurs et leur schéma linguistique sur différents types de corpus : une première étude a été menée sur quatre types de corpus pour les marqueurs de la base générique ; elle peut être approfondie en évaluant plus précisément l'impact de l'opérationnalisation du marqueur ;
- envisager une analyse syntaxique préalable, et un travail sur un corpus étiqueté ; ce type d'analyseur étant disponible, les modalités de définition et d'opérationnalisation des marqueurs pourraient être élargies ;
- intégrer ce type de logiciel à un environnement de modélisation de connaissances à partir de texte ; pour l'instant, les principes de Caméléon seront repris en un module dans l'environnement Terminae de construction de terminologies et d'ontologies (Biébow et Szulman, 2000).

## Références bibliographiques

- Assadi, H., Bourigault, D. (1996), « Acquisition de connaissances à partir de textes : Outils informatiques et éléments méthodologiques », in *Actes du dixième congrès Reconnaissance de Formes et Intelligence Artificielle (RFIA '96)*, Rennes, pp. 505-514.
- Aussenac-Gilles, N. & Condamines, A. (1999), « Bases de connaissances terminologiques : enjeux pour la consultation documentaire », in J. Maniez et W. Mustapha El Hadi (éds), *Organisation des connaissances*

*en vue de leur intégration dans les systèmes de représentation et de recherche d'information*, Villeneuve d'Asq, Univ. Charles de Gaulle, (UL3 travaux et recherches), pp. 71-88.

- Aussenac-Gilles, N., Biébow, B. & Szulman, S. (2000), « Revisiting Ontology Design : a methodology based on corpus analysis », in R. Dieng & O. Corby (eds), *Proceedings of the 12<sup>th</sup> International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000)*, Juan-les-Pins, October 2000, Springer Verlag. LNAI 1937, pp. 172-188.
- Biébow, B. & Szulman, S. (1997), « Méthodologie de création d'un noyau de base de connaissances en logique terminologique à partir de textes », in *Actes des deuxièmes rencontres Terminologie et Intelligence Artificielle. TIA'97*, Toulouse, Ed. ERSS et Université Toulouse-Le Mirail, pp. 69-84.
- Biébow, B. & Szulman, S. (2000), « Terminae : une approche terminologique pour la construction d'ontologies du domaine à partir de textes », in *Actes de RFIA 2000, Reconnaissances des Formes et Intelligence Artificielle*, Paris.
- Borillo, A. (1996), « Exploration automatisée de textes de spécialité : repérage et identification automatique de la relation lexicale d'hyponymie », in *Linx*, 34-35, pp. 113-121.
- Bourigault, D. (1996), « Lexter, a Natural Language Processing Tool for Terminology Extraction », in *Proceedings of Euralex'96*, Göteborg University, Department of Swedish, pp. 771-779.
- Bourigault, D. & Jacquemin, C. (2000), « Construction de ressources terminologiques », in J.-M. Pierrel (éd), *Ingénierie des langues*, Traité I2C, Paris, Hermes.
- Brachman, R. (1977), « What's in a concept : structural foundations of semantic networks », in *Intern. Journal of Man Machine Studies*, 9, pp. 127-152.
- Brachman, R. (1979), « On the epistemological status of semantic networks », in N. Findler (ed), *Associative Networks : representation and use of knowledge by computer*, Academic Press, pp. 3-50.
- Condamines, A. & Rebeyrolle, J. (1998), « CTKB : a corpus based approach to a Terminological Knowledge Base », in *Proceedings of Computerm'98, 1<sup>st</sup> workshop on Computational Terminology*, pp. 29-35.
- Condamines, A. & Rebeyrolle, J. (2000a), « Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode », in J. Charlet, M. Zacklad, G. Kassel & D. Bourigault, (éds), *Ingénierie des Connaissances, évolutions récentes et nouveaux défis*, Paris, Eyrolles, pp. 225-241.
- Condamines, A. & Rebeyrolle, J. (2000b), « Searching for and Identifying Conceptual Relationships via a corpus-based approach to a

- Terminological Knowledge Base (CTKB) : method and results », in D. Bourigault, M.-C. L'Homme & C. Jacquemin (eds), *Recent Advances in Computational Terminology*, John Benjamins.
- David, S. & Plante, P. (1990), *Termino version 1.0*. Rapport de recherche du Service d'Analyse de Textes par Ordinateurs (ATO), Université du Québec à Montréal. <http://www.ling.uqam.ca/nomino>
- Daoust, F. (1996), *SATO (Système d'analyse de texte par ordinateur), Version 4.0, Manuel de référence*, Service d'analyse de texte par ordinateur (ATO), Université du Québec à Montréal, 256 p. <http://www.ling.uqam.ca/sato>
- Garcia, D. (1998), *Analyse automatique des textes pour l'organisation causale des actions, Réalisation du système Coatis*. Thèse d'informatique, Université Paris IV.
- Jackiewicz, A. (1996), « L'expression lexicale de la relation d'ingrédience (partie-tout) ». *Faits de Langues* 7, pp 53-62, Orphys.
- Jouis, C. (1993), *Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse de textes. Réalisation d'un prototype : le système Seek*. Thèse en informatique, EHESS, Paris.
- Kayser, D. (1997), *La représentation des connaissances*, Paris, Hermès.
- Meyer, I. (2000), « Extracting Knowledge-rich Contexts for Terminography : A Conceptual and methodological Framework », in D. Bourigault, M.-C. L'homme & C. Jacquemin (eds), *Recent Advances in Computational Terminology*, John Benjamins.
- Morin, E., (1998), « Prométhée : un outil d'aide à l'acquisition de relations sémantiques entre termes », in *Actes TALN*, Paris, pp. 172-181.
- Newell, A. (1982), « The Knowledge Level », in *Artificial Intelligence*, 18, pp. 87-127.
- Otman, G. (1996), « Le traitement automatique de la relation partie-tout en terminologie », in *Faits de Langue*, 7, mars 1996.
- Rebeyrolle, J. (2000), *Forme et fonction de la définition dans le discours*. Thèse de doctorat de l'Université de Toulouse 2.
- Rebeyrolle, J. & Tanguy, L. (2001), « Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires ». in *Cahiers de grammaire*. Ce numéro.
- Rousselot, F., Frath, P. & Oueslati, R. (1996), « Extracting Concepts and relations from corpora », in *Proceedings ECAI'96, 12<sup>th</sup> European Conference on Artificial Intelligence*.
- Séguéla, P. (1999), « Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés », in *Actes de TIA'99 (Terminologie et Intelligence Artificielle), Terminologies Nouvelles* 19, pp. 52-60.
- Séguéla, P. & Aussenac-Gilles, N. (1999), « Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine », in *Actes de IC'99 (Ingénierie des Connaissances)*, Paris, pp. 79-88.