Généralisation de structures prédicatives. Application à l'analyse de l'information

Nicolas Capponi

Département de formation doctorale en informatique
UFR STMIA

École doctorale IAE + M

Généralisation de structures prédicatives. Application à l'analyse de l'information

THÈSE

présentée et soutenue publiquement le 8 janvier 1999

pour l'obtention du

Doctorat de l'université Henri Poincaré – Nancy 1

(spécialité informatique)

par

Nicolas Capponi

Composition du jury

Président :

Rapporteurs : Daniel Kayser, Professeur à l'Université Paris 13
Marie-Christine Rousset, Professeur à l'Université Paris-Sud
Didier Galmiche, Maître de Conférences à l'Université Henri Poincaré, Nancy

Examinateurs : Jean-Marie Pierrel, Professeur à l'Université Henri Poincaré, Nancy
Amedeo Napoli, Chargé de Recherche au CNRS, Nancy
Yannick Toussaint, Chargé de Recherche à l"INRIA Lorraine, Nancy

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Table des matières

Table des figures

1

Motivations et objectifs

Nous présentons dans ce mémoire une méthode originale de structuration de structures prédicatives par généralisation. Nous appliquons cette méthode à l'analyse de textes scientifiques dans le but d'améliorer une chaîne de traitement existante, qui est limitée par le type d'informations qu'elle manipule : des termes et des associations entre termes. L'utilisation de structures prédicatives permet de proposer une analyse plus fine, et leur structuration par généralisation permet de disposer d'une vue synthétique du contenu informationnel des textes, contribuant à améliorer le processus d'analyse.

L'analyse de l'information scientifique et technique

Nous nous intéressons au traitement de l'information à caractère scientifique et technique, dans le cadre de l'analyse de l'information. Dans un contexte d'augmentation très rapide de l'information sous une forme électronique (banques de données documentaires, documentation technique, ...), il devient nécessaire de disposer de méthodes informatiques permettant de bien maîtriser l'accès à cette information.
L'analyse de l'information vise à caractériser le contenu d'un ensemble de textes afin d'en donner une vision globale : il s'agit de permettre une vue synthétique d'une masse de documents, en faisant émerger le contenu informationnel au moyen de méthodes et d'outils informatiques. In fine, l'analyse est effectuée par un expert d'un domaine de spécialité donné, qui s'appuie sur le contenu extrait des documents pour en réaliser l'interprétation. L'objectif final peut être la constitution d'un dossier de veille technologique, d'un rapport de tendance à destination de décideurs ou plus simplement l'établissement d'un état de l'art sur un sujet particulier.

Concrètement, les méthodes d'analyse consistent à extraire des informations d'un corpus de documents afin de les structurer et de les présenter à l'utilisateur qui pourra les interpréter. Ces informations prennent donc la forme d'unités langagières : mots et groupes de mots[1] qui renvoient aux connaissances véhiculées par les textes. L'identification des relations entre ces unités par le biais de la co-occurrence des termes, notamment au moyen de méthodes statistiques, permet d'opérer une structuration des unités d'information et de fournir ainsi une « image » organisée du contenu informationnel des textes, à partir de laquelle intervient le spécialiste du domaine. Toutefois, cette « image », qui prend la forme

1. Dans un domaine de spécialité, il s'agit de termes simples (un mot) et complexes (plusieurs mots), qui renvoient à des concepts du domaine.

de classes de termes (ensemble de termes et de relations entre ces termes) organisées au sein d'une carte thématique, nécessite un travail important de la part de l'analyste pour aboutir à la caractérisation finale de l'ensemble des documents.

Les structures prédicatives pour accéder aux informations des textes

Bien qu'il soit à notre sens irréaliste de prétendre automatiser complètement un tel processus, nous défendons dans cette thèse une approche plus exhaustive, proposée dans le cadre du projet ILC (Infométrie, Langage et Connaissances)[2] : nous faisons l'hypothèse qu'une analyse plus fine et plus complète, allégeant le travail de l'analyste, peut être obtenue en recherchant dans les textes des informations plus élaborées : les structures prédicatives. Les structures prédicatives mettent en jeu plusieurs unités d'information représentées par des unités linguistiques de la forme prédicat-arguments. Classiquement, un prédicat représente une action ou un événement lié au domaine, et les arguments précisent les objets impliqués dans l'action ou l'événement. Une structure prédicative permet ainsi de rendre compte de manière plus fine du contenu informatif des textes que des associations de co-occurrence entre termes, qui sont par nature peu précises.
Nous pouvons illustrer ceci par un exemple simple. Dans le domaine de l'agriculture, l'analyse d'un corpus de textes peut faire émerger les relations de co-occurrence suivantes entre termes :

dosage et *amine*
dosage et *chromatographie*
amine et *chromatographie*

L'analyste, pour en savoir plus sur le contexte de ces associations, doit avoir recours aux textes où apparaissent conjointement ces termes. L'utilisation de structures prédicatives permet d'obtenir une information plus riche, sans nécessiter un retour aux textes. Ainsi, la structure prédicative suivante :

dosage(objet : amine, moyen : chromatographie)

permet de préciser comment les termes *dosage*, *amine* et *chromatographie* sont liés. *Objet* et *moyen* sont des rôles thématiques qui précisent le rapport entre le prédicat et l'argument. Ainsi, le dosage porte sur des substances, les amines, et le moyen utilisé est une technique d'analyse appelée chromatographie.

Nos objectifs

Dans cette thèse, nous montrons que l'utilisation de structures prédicatives peut améliorer de façon sensible le processus de l'analyse de l'information. Cependant leur utilisation repose sur une structuration qui présente une vision synthétique du contenu informationnel : l'information collectée doit être organisée afin d'être exploitée efficacement. Notre travail s'est donc décomposé en deux objectifs :

- proposer une méthode de structuration par généralisation de structures prédicatives extraites des textes, qui prenne en compte les spécificités de la nature linguistique de ces structures.

2. Le projet ILC est développé dans le cadre d'une collaboration entre l'équipe RFIA du LORIA-INRIA Lorraine et le Programme de Recherche Infométrie de l'INIST.

- montrer que la méthode proposée trouve une application naturelle en analyse de l'information, en l'intégrant dans une chaîne de traitements de textes qui repose sur une approche terminologique et qui permet l'analyse du contenu informationnel d'un ensemble de documents.

Le coeur de notre travail est la proposition d'une méthode originale de généralisation de structures prédicatives. La généralisation de structures prédicatives n'a pas fait l'objet, à notre connaissance, de travaux spécifiques. Toutefois, elle s'inscrit, au moins partiellement, dans les recherche en apprentissage, et plus particulièrement en *classification conceptuelle*, dont l'objectif est la construction automatique de hiérarchies à partir d'un ensemble d'objets.

L'application de notre méthode de généralisation à l'analyse de l'information représente le deuxième volet de notre travail : elle nous permet d'évaluer la méthode et d'étudier l'aspect linguistique de façon plus approfondie.

Plan du mémoire

Notre mémoire se compose de deux grandes parties.
La première partie présente la méthode de généralisation de structures prédicatives que nous avons élaborée.
La deuxième partie présente une application de notre méthode à l'analyse de l'information, dont l'objectif est de synthétiser l'information contenue dans des documents textuels.
Nous nous sommes peu intéressés à la manière dont on obtient ces structures prédicatives car ce n'était pas notre objectif prioritaire. Le lecteur ne devra donc pas s'attendre à trouver une méthode d'extraction de structures prédicatives, mais plutôt une méthode de structuration et de synthèse de ces structures.
La première partie est focalisée sur la généralisation, et met l'accent sur la partie logique de notre travail. L'aspect linguistique n'est qu'abordé que pour justifier certains choix réalisés. Aussi certains aspects très importants, notamment le lien entre les unités linguistiques (les termes) et les concepts ne sont pas discutés dans cette partie, mais sont abordés plus tard. Dans la deuxième partie, nous nous intéressons de manière beaucoup plus approfondie aux aspects linguistiques de ce travail. Les méthodes et outils permettant la mise en oeuvre d'un processus automatique d'analyse de l'information sont discutés et l'articulation entre l'aspect linguistique et l'aspect logique est traité de façon plus complète.

La première partie, consacrée à la généralisation, est constituée de trois chapitres (chapitres 2, 3 et 4).
Le chapitre 2 fait l'inventaire des méthodes de généralisation existantes, principalement issues du domaine de la classification conceptuelle. Nous distinguons les deux approches principales utilisées en classification conceptuelle, et montrons l'importance du formalisme de représentation de connaissances utilisé. L'étude des différents travaux nous permet de choisir un formalisme de représentation de connaissances adéquat, les logiques de descriptions.
Dans le chapitre 3, nous présentons en détail les connaissances que nous sommes amenés à généraliser, à savoir les structures prédicatives, et montrons comment nous représentons ces connaissances avec le formalisme de logique de descriptions utilisé, CLASSIC.
Le chapitre 4 constitue le coeur de notre travail : nous y présentons notre méthode de généralisation avec la logique de descriptions CLASSIC. Les heuristiques utilisées, ainsi que les

algorithmes correspondants sont détaillés. La complexité de ce processus de généralisation est exposée.

La deuxième partie, qui montre comment appliquer notre méthode de généralisation à l'analyse de l'information, est constituée de deux chapitres (chapitres 5 et 6).
Le cinquième chapitre de notre thèse présente en détail le processus d'analyse de l'information. Nous étudions les méthodes et outils informatiques constitutifs d'une chaîne de traitement de documents textuels pour l'analyse. Nous montrons en particulier la nécessité d'une approche terminologique, et proposons l'utilisation de structures prédicatives comme moyen d'améliorer la finesse de l'analyse.

Le sixième chapitre présente une première évaluation du processus de généralisation dans le cadre de l'analyse de l'information. Au travers d'une expérimentation sur un corpus de résumés du domaine de l'agriculture, nous positionnons notre processus dans la chaîne de traitement et montrons comment il permet de proposer une vue synthétique des textes, permettant un accès plus efficace et plus pertinent à l'information qu'ils contiennent.

Nous terminons ce mémoire par un résumé des principaux résultats obtenus, complété par un aperçu des perspectives offertes par notre travail.

2

Apports et limites des travaux en classification conceptuelle

Nous présentons dans ce chapitre une étude critique des travaux en classification conceptuelle qui sont susceptibles d'être mis à contribution pour notre objectif de généralisation de structures prédicatives. Nous commençons par donner une ensemble de définitions des principales notions que nous utiliserons (section 2.1). Puis nous présentons une vue d'ensemble des travaux en classification conceptuelle (section 2.2). Nous montrons ensuite les limites des approches classiques (section 2.3), et nous intéressons à des approches utilisant différents formalismes de représentation des connaissances : les treillis de concepts, les graphes conceptuels et les logiques de descriptions (section 2.4 à 2.6). Sur la base des résultats de cette étude, nous concluons par le choix d'un formalisme qui sera utilisé comme cadre pour la généralisation des structures prédicatives (section 2.7).

2.1 Définitions des principales notions utilisées

Les relations entre les notions de *généralisation*, de *classification*, de *hiérarchie* et de *concept*, ainsi que quelques notions annexes, sont présentées ci-après à travers un ensemble de définitions qui permettront au lecteur de mieux appréhender notre problématique et la présentation de notre travail. Ces définitions n'ont pas l'ambition d'être normatives, mais plutôt de situer notre point de vue sur des termes souvent flous et ambigus.

Domaine, concept et objet

En intelligence artificielle, on s'intéresse très souvent à la représentation d'un *domaine de référence*, qui est une modélisation du monde ou d'une partie du monde. Un domaine est décrit à l'aide d'*objets* individuels, et de *concepts*, regroupant par abstraction différents objets.

Un **concept** peut s'appréhender de deux manières complémentaires :

- une vision ensembliste permet de considérer le domaine de référence comme un ensemble, les objets sont alors les éléments de cet ensemble, et les concepts des

sous-ensembles[3]. Les objets qui appartiennent au sous-ensemble correspondant à un concept sont des *instances* de ce concept. L'ensemble des instances d'un concept est l'**extension** de ce concept.

- alternativement, un concept peut être perçu comme un ensemble de *caractéristiques* ou *propriétés*. Une instance de concept est alors un objet qui satisfait les propriétés d'un concept. L'ensemble des propriétés d'un concept constitue l' **intension** de ce concept.

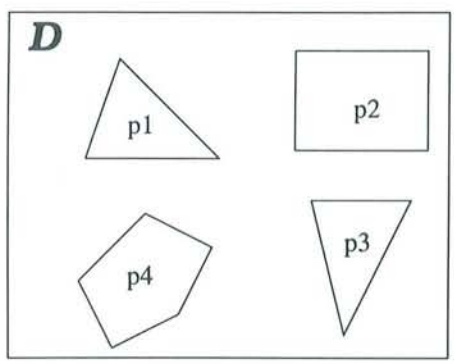

FIG. 2.1 – *Un domaine de référence, $\mathcal{D}$, composé de quatres objets géométriques*

Par exemple, soit un domaine $\mathcal{D}$ comprenant quatre objets p_1, p_2, p_3, p_4 (voir figure 2.1). Le concept de *triangle* a pour extension l'ensemble $\{p_1, p_3\}$. Une définition en intension du concept de triangle peut être : *polygone ayant trois cotés* ou bien alternativement *polygone ayant trois angles.*

Classification

Le terme **classification**, même en se restreignant au domaine de l'intelligence artificielle, possède de nombreux sens. Il peut en effet désigner :

1 un *processus inductif* qui permet d'aboutir à cette structure,

2 un *processus déductif* qui permet d'identifier la classe d'appartenance d'un objet,

3 un *processus déductif* qui permet d'insérer un nouvel objet ou une nouvelle classe dans cette structure,

4 une *structure* qui organise des objets en classes (ou concepts),

Dans ce mémoire, nous serons amenés à discuter de ces différents sens. Par défaut, le terme *classification* sera utilisé pour désigner un processus inductif [sens 1][4]. Pour désigner la recherche de la classe d'un objet [sens 2], nous parlerons d'**identification**. Le sens [3] est utilisé dans la section où nous présentons le formalisme des logiques de descriptions. Nous conserverons le terme *classification*, en précisant qu'il s'agit d'un mécanisme déductif. Enfin, pour désigner une structure, nous utiliserons les termes décrits ci-dessous (paragraphe suivant).

3. Dans cette vision ensembliste, les concepts sont également appelés *classes.*

4. Le processus inductif est parfois désigné par le terme **catégorisation**. Ce terme a été choisi par plusieurs auteurs [Napoli 96] [Euzenat 94]. D'autres préfèrent réserver ce terme pour désigner, en psychologie cognitive, l'opération *mentale* de regroupement d'objets semblables [Bournaud 96, page 18].

Structures classificatoires et hiérarchies

Une *structure classificatoire* peut prendre diverses formes. Nous nous intéresserons essentiellement aux structures hiérarchiques, et nous utiliserons le terme *hiérarchie* pour les désigner[5]. Nous définissons une **hiérarchie** $\mathcal{H} = (\mathcal{X}, \sqsubseteq, \omega)$ comme un graphe orienté sans circuits où :

- $\mathcal{X}$ est un ensemble de classes,
- $\sqsubseteq$ est une relation d'ordre partiel[6], que l'on appellera *relation de spécialisation* ou *relation de subsomption*,
- ω est l'élément maximal de $\mathcal{X}$ suivant $\sqsubseteq$, existe toujours, et est appelé la *racine* de la hiérarchie.

Les classes représentent les sommets du graphe, et les arcs correspondent aux relations entre les classes. Si l'on considère que l'ensemble $\mathcal{X}$ représente des concepts plutôt que de simples classes, on peut utiliser le terme *hiérarchie conceptuelle.*
Étant donné un élément C de la hiérarchie $\mathcal{H}$, ses *pères* représentent les antécédents directs de C dans $\mathcal{H}$; ses *fils* représentent les descendants directs de C.
Lorsque deux éléments C et D de la hiérarchie $\mathcal{H}$ vérifient $C \sqsubseteq D$, on dit que D *subsume* (est plus général que) C. C est le *subsumé*, et D est le *subsumant.*

Une *hiérarchie stricte* est une hiérarchie $\mathcal{H}$ où toutes les classes sont disjointes, c'est-à-dire où chaque élément C de $\mathcal{H}$ possède un seul père.
Un *treillis* est une hiérarchie $\mathcal{H}$ telle que tout couple d'éléments (C, D) possède un maximum $C \wedge D$ et un minimum $C \vee D$ uniques.

Les deux autres principaux type de structures classificatoires sont la *partition*, simple répartition d'un domaine en classes disjointes, et le *recouvrement*, répartition d'un domaine en classes non nécessairement disjointes [Decaestecker 93]. La figure 2.2 illustre les différents types de structures classificatoires présentées.

Généralisation

Étant donné une hiérarchie $\mathcal{H}$, la **généralisation** d'un ensemble X d'objets ou de concepts[7] est une opération qui consiste à calculer un concept g qui soit plus général (au sens de la relation de subsomption) que chacun des éléments de X. Le concept g est appelé un *concept généralisation* de X[8].
Parmi les concepts généralisations possibles de X, l'ensemble des *plus petites généralisations communes* sont ceux qui ne subsument pas un autre concept généralisation de X. Autrement dit, s'il n'existe pas de concept généralisation c de X tel que $c \sqsubseteq g$, alors g fait partie des plus petites généralisations communes de X.

Par extension, la généralisation d'un ensemble d'objets (ou de concepts) est un processus qui consiste à trouver un ensemble de concepts généralisations qui permette une organisation synthétique de l'ensemble d'objets (ou de concepts) en hiérarchie.

5. Le terme *taxinomie* est aussi utilisé en intelligence artificielle.

6. On se limite parfois à un pré-ordre, c'est-à-dire à une relation transitive, réflexive mais non nécessairement antisymétrique [Leclère 96].

7. X étant éventuellement réduit au singleton.

8. Lorsqu'il n'y a pas d'ambiguïté, nous utilisons simplement le terme *généralisation.*

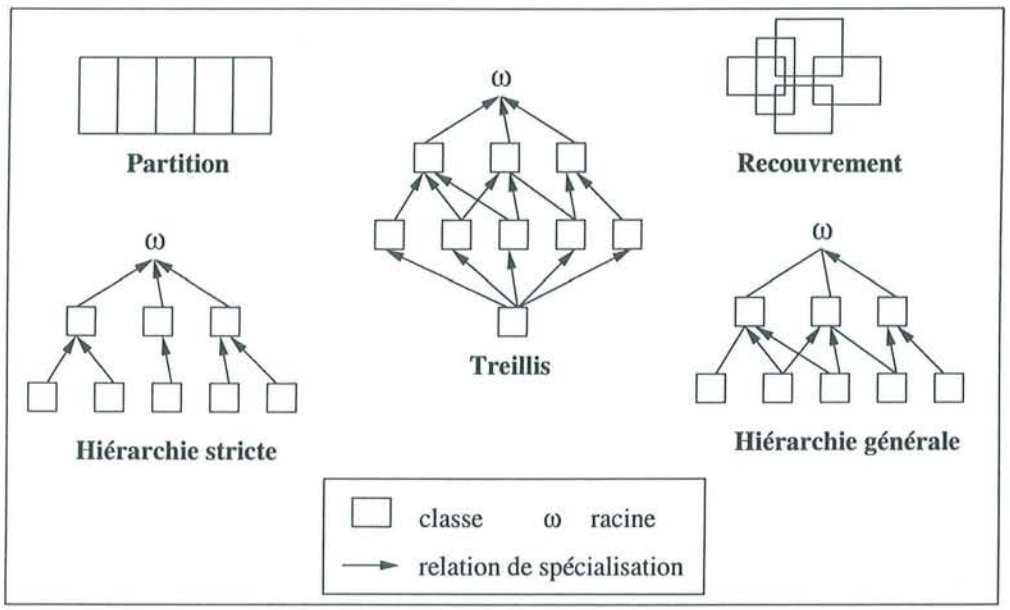

FIG. 2.2 – *Les principaux types de structures classificatoires*

Notre objectif de généralisation de structures prédicatives peut ainsi être défini de manière plus précise.
Étant donnés :

- un ensemble de structures prédicatives représentées par des concepts,
- une organisation hiérarchique des éléments (concepts) constituant les structures prédicatives,

Il s'agit de :

- trouver un ensemble de concepts généralisations synthétisant les structures prédicatives[9],
- trouver une organisation hiérarchique de ces concepts.

Nous pouvons à présent nous intéresser à l'étude de l'état de l'art en classification conceptuelle.

2.2 Vue d'ensemble des travaux en classification conceptuelle

Nous présentons une vue d'ensemble de la classification conceptuelle, qui montre que l'objectif principal de ce domaine de recherche est la construction automatique de structures classificatoires, et principalement de hiérarchies. Nous présentons ensuite les deux grands types d'approches qui ont été utilisées, et montrons l'importance du formalisme de représentation des connaissances pour la classification conceptuelle.

9. Les concepts généralisations représentent également des structures prédicatives. Pour être plus précis, lorsque nous employons le terme « structure prédicative », il s'agit ici du « concept représentant une structure prédicative ». Nous discutons ce point plus en détail au chapitre 3.

2.2.1 Un objectif : construire automatiquement des hiérarchies

La **classification conceptuelle**[10] est un champ de l'intelligence artificielle concerné par la construction automatique de structures classificatoires, et plus particulièrement de hiérarchies. Initialement, les travaux de classification automatique se sont orientés vers des approches statistiques, qui sont maintenant approfondies dans le cadre de l'analyse de données [Saporta 90] [Celeux 95] comme le montre Ketterlin qui dresse un panorama de l'activité [Ketterlin 95, pages 7–14]. La différence majeure entre les deux approches est l'introduction de la notion de *concept* par les chercheurs en intelligence artificielle, qui permet de prendre en compte des données symboliques et de fournir ainsi une *description* des classes qui regroupent les objets.

La classification conceptuelle est une forme d'*apprentissage non supervisé*, dans la mesure où les objets à regrouper en concepts n'ont pas été classés a priori. Elle s'oppose à l'*apprentissage supervisé*[11] dont l'objectif est de trouver les concepts permettant d'organiser un ensemble d'objets préalablement divisés en exemples et contre-exemples.

Une définition précise de la classification conceptuelle à été donnée par Michalski et Stepp [Michalski 83] :

> Étant donné un ensemble d'objets et leur descriptions associées, trouver :
> - un ensemble de classes regroupant ces objets, et maximisant les similitudes au sein d'une classe et les différences entre classes distinctes,
> - une définition intensionnelle de chacune des classes,
> - une organisation hiérarchique des classes.

Une distinction est faite selon que le processus est incrémental ou non. Le processus est dit *incrémental* lorsqu'il traite les objets de manière séquentielle : étant donnée la hiérarchie $\mathcal{H}$, le processus définit une opération d'insertion d'un nouvel objet dans la hiérarchie, qui modifie celle-ci en conséquence. Un processus *non incrémental* considère l'ensemble des objets d'un bloc, et retourne la hiérarchie correspondante. La classification conceptuelle incrémentale est connue également sous le terme *formation de concept*[12].

Plusieurs utilisations des hiérarchies

Les hiérarchies construites automatiquement sont destinées à différents objectifs, dont les plus courants sont [Bournaud 96] [Godin 95] :

- la prédiction de valeurs de caractéristiques inconnues pour de nouveaux objets,
- l'aide à la découverte scientifique,
- l'organisation de connaissances,

Etant donné un nouvel individu X ne participant pas à la hiérarchie de concepts, la *prédiction de valeurs* de caractéristiques inconnues consiste à déduire la valeur d'une propriété a concernant cet individu. Cette propriété a peut être simplement la classe à laquelle appartient X, et le problème est alors celui de l'*identification* du nouvel objet X.

10. Traduction de l'expression *conceptual clustering*. On trouve également dans la littérature les expressions *regroupement conceptuel* et *catégorisation conceptuelle*.

11. Le terme *apprentissage de concept* est également utilisé.

12. Traduction de *concept formation*.

Comme le fait remarquer I. Bournaud [Bournaud 96], l'avantage de ce type d'utilisation est la facilité de l'évaluation du processus : l'ensemble des objets du domaine est divisé en deux ensembles, l'un étant utilisé pour l'apprentissage, et l'autre pour tester automatiquement la prédictivité de la structure classificatoire construite.

L'*aide à la découverte scientifique* doit permettre de formuler une théorie ou une loi empirique, en utilisant la structure classificatoire obtenue sur un grand nombres d'objets. L'accent est mis sur la comparaison entre différentes structures obtenues afin de faire émerger la théorie ou la loi.

L'*organisation de connaissances* consiste à trouver une structure classificatoire qui soit pertinente pour rendre compte des similarités et des différences structurelles entre des objets. Une structure hiérarchique est bien adaptée à l'organisation des connaissances, et est utilisée par de nombreuses disciplines (biologie, psychologie, représentation par objets, réseaux sémantiques, ...).

Notre objectif nous situe naturellement dans cette dernière catégorie, puisque nous considérons la généralisation de structures prédicatives comme un moyen synthétique d'organiser des informations extraites de textes. Les méthodes proposées en classification conceptuelle diffèrent selon l'objectif visé, et nous nous focaliseront donc essentiellement sur celles qui favorisent l'organisation de connaissances.

2.2.2 Les deux types d'approches en classification conceptuelle

A la suite de Bournaud [Bournaud 96], nous distinguons deux grandes approches en classification conceptuelle : l'*approche classique* et l'*approche basée sur un Espace de Connaissances*. Ces deux approches sont schématisées sur les figures 2.3 et 2.4.

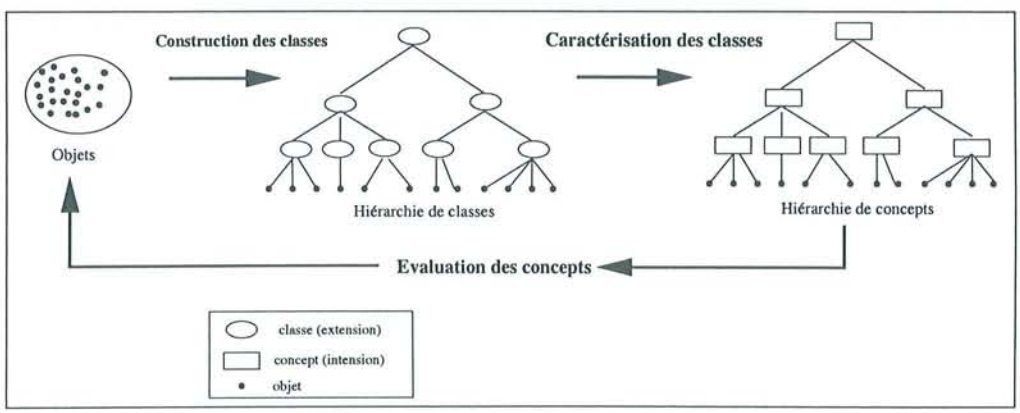

FIG. 2.3 – *Approche « classique » en classification conceptuelle d'après [Bournaud 96]*

L'approche classique

L'approche classique peut se caractériser par un découpage en trois procédures principales [Bisson 92] [Bournaud 96] :

- une procédure de construction de classes sous forme hiérarchique,

- une procédure de caractérisation des classes, qui fournit une description intensionnelle des classes,
- une procédure d'évaluation de la hiérarchie, qui permet de juger de la qualité du résultat (notamment la capacité de prédiction).

La *procédure de construction* nécessite de définir des critères de regroupement des objets et des classes. Étant donné un ensemble X constitué de n objets, il faut trouver les sous-ensembles C_i qui vont constituer les classes. Le nombre de sous-ensembles possibles est le cardinal de $\mathcal{P}(X)$, l'ensemble des parties d'un ensemble, qui grandit de manière exponentielle et ne permet pas d'envisager une solution exhaustive. Même en se restreignant à un ensemble de partitions (c'est-à-dire que les classes proposées sont exclusives), il existe plus de 2^n solutions, qu'il n'est pas pensable de construire pour aboutir à un résultat exploitable pour de grandes valeurs de n.

Les critères de regroupement permettent de ne retenir qu'un ensemble restreint de solutions, correspondant à la « meilleure » classification possible. Ces critères peuvent être basés sur :

- une estimation globale : la plus utilisée est celle qui essaye de maximiser la similarité au sein d'une classe et de minimiser la différence entre classes distinctes en s'inspirant de la notion de *category utility* issue des travaux en psychologie cognitive (elle est présentée section 2.3),
- une estimation locale : une *mesure de similarité* entre objets permet de sélectionner les objets à regrouper.

Le *processus de caractérisation*, qui permet la description des concepts, est souvent basé sur les algorithmes de l'apprentissage supervisé, utilisant des exemples et contre-exemples de concepts. Il suffit en effet de considérer les objets regroupés au sein d'une classe comme les exemples et les autres objets comme des contre-exemples.

La *procédure d'évaluation* est souvent très liée au critère de regroupement, utilisant ce dernier comme mesure.

Nous appelons cette approche *classique* car elle est représentative des premiers travaux en classification conceptuelle, et a été poursuivie par une grande majorité des chercheurs du domaine. Elle est cependant essentiellement concernée par la prédiction de valeurs. Nous détaillons un des représentants les plus connus de cette approche, COBWEB, dans la section 2.3 et montrons qu'elle n'est pas très adaptée à notre objectif.

L'approche de type Espace de Connaissances

L'alternative à l'approche classique s'est développée au cours des dernières années, notamment à partir des travaux sur les treillis de concepts [Godin 95]. Elle se distingue de l'approche classique en offrant notamment :

- la prise en compte de connaissances sur le domaine, le plus souvent sous la forme d'une hiérarchie de concepts,
- un processus de généralisation des objets guidé par les connaissances du domaine, permettant de construire un Espace de Connaissances (EC) [13] . L'EC est une hiérarchie

13. Nous utilisons l'appellation *Espace de Connaissances* proposée par Mineau [Mineau 90]. Bournaud [Bournaud 96] lui préfère le terme *Espace de Généralisations*, mais il s'agit de la même structure.

de concepts « objective », dans le sens où tous les concepts généralisation exprimables dans le langage de représentation utilisé sont calculés,

- un processus d'extraction opérant sur l'EC, qui permet d'obtenir des hiérarchies élaguées selon divers modes. Une hiérarchie peut par exemple représenter un *point de vue* particulier sur les objets[14].

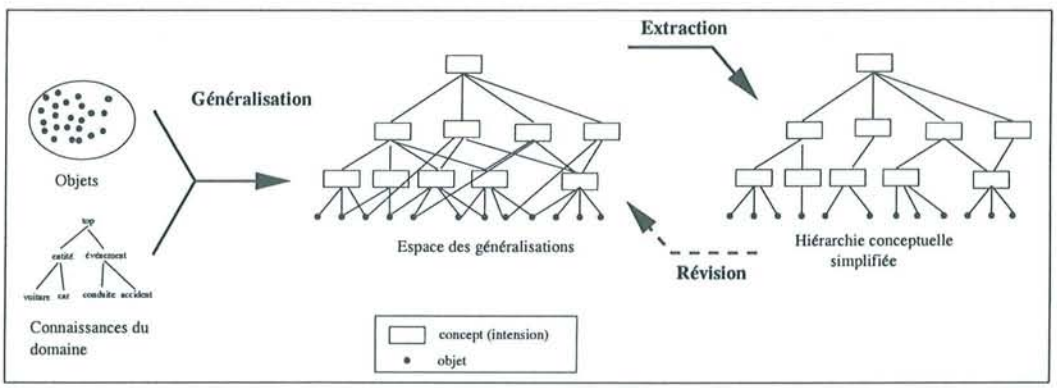

FIG. 2.4 – *Approche « Espace de Connaissances » en classification conceptuelle d'après [Bournaud 96]*

Nous présentons les diverses orientations de cette approche dans les sections 2.4 et 2.5, en montrant les différents avantages offerts par rapport à l'approche classique. Nous verrons qu'elles diffèrent essentiellement par l'utilisation de différents formalismes de représentation. Ce point important est discuté dans la section suivante.

2.2.3 Importance de la représentation des connaissances pour la classication conceptuelle

La question de la *représentation des connaissances* est essentielle car elle conditionne le type d'objets et de concepts que l'on peut manipuler. La représentation s'effectue par l'intermédiaire d'un *formalisme de représentation des connaissances*, conçu pour expliciter la connaissance relative à un problème, afin de pouvoir effectuer des raisonnements sur celle-ci grâce à des mécanismes d'inférences. En général, le formalisme utilisé pour décrire les concepts est le même que celui utilisé pour décrire les objets. Le choix d'un formalisme repose sur un ensemble de critères, dont les principaux sont [Nebel 90a] [Cercone 87] :

- la puissance expressive du formalisme,
- la lisibilité et la clarté d'expression du formalisme,
- la définition d'une sémantique formelle, afin d'expliciter de manière non ambiguë le sens des expressions du langage,
- la disponibilité de mécanismes d'inférences déductifs, qui soient corrects et complets conformément à la sémantique formelle du langage, et de complexité calculatoire réduite,
- la disponibilité de mécanismes d'inférences inductifs, de complexité calculatoire réduite.

14. Pour une discussion sur la classification déductive avec points de vue, voir [Mariño 93].

La réunion de tous ces critères est bien évidemment hors de portée : il existe un compromis important entre l'expressivité d'un langage et l'efficacité des inférences qui lui sont associées, qu'elles soient déductives ou inductives. Il faut donc considérer un langage qui soit suffisamment expressif pour représenter les données d'un problème et suffisamment contraint pour ne pas conduire à des algorithmes inutilisables. En intelligence artificielle, le langage de référence reste la *logique des prédicats*, bien maîtrisée pratiquement et théoriquement. Cependant la logique des prédicats est un langage trop expressif pour pouvoir être utilisé avec un processus inductif [Kietz 94] [Ventos 95] : de nombreux travaux se sont ainsi consacrés à la recherche de restrictions de la logique des prédicats présentant des propriétés inductives « satisfaisantes » [Saitta 96].

On distingue généralement de ce point de vue les *langages propositionnels* et les *langages relationnels* [Saitta 96]. Un langage propositionnel décrit un concept comme la conjonction d'un ensemble de paires attribut/valeur (A_i, V_i) où les V_i sont des valeurs nominales. Un langage relationnel est analogue à la logique des prédicats, mais peut être restreint de diverses manières, par exemple en n'autorisant que des prédicats unaires ou binaires au lieu de prédicats n-aires. Un formalisme intermédiaire intéressant est ainsi l'extension du formalisme attribut/valeur à des valeurs pouvant être des objets structurés [Ketterlin 95] [Thompson 91]. Dans le même ordre d'idées, les formalismes de représentation par objets, ou des formalismes apparentés, comme ceux des *logiques de descriptions*, des *graphes conceptuels* et des *treillis de Galois* ont été utilisés en classification conceptuelle. Ils sont bien adaptés aux approches de type *Espace de Connaissances*. Nous nous focalisons par la suite sur ces différents formalismes, en montrant les avantages et limites liés à chacun d'entre eux pour une approche de type EC (sections 2.4, 2.5 et 2.6).

Cependant, certaines approches en classification conceptuelle ont préféré utiliser une définition probabiliste des concepts. Un *concept probabiliste* est décrit par un ensemble d'attributs, où chaque attribut se voit associer une distribution de probabilités, qui représente la fréquence d'apparition des différentes valeurs de cet attribut pour les objets qui sont instances du concept. Par exemple, un concept C ayant un attribut *couleur* et un attribut *taille* peut posséder les distribution suivantes :

$((jaune, 0.2)(rouge, 0.4)(bleu, 0.4))$ et
$((petit, 0.5)(moyen, 0.4)(grand, 0.1))$

A un niveau supérieur, une probabilité d'apparition est associée à chaque concept. Cette représentation est utilisée par le système COBWEB et certains de ses descendants qui représentent l'approche classique de la classification conceptuelle, et sont présentés à la section suivante.

Nous pouvons enfin mentionner un autre formalisme : la programmation logique inductive, qui a fait l'objet de nombreux travaux pour rechercher un langage offrant un bon compromis entre expressivité et complexité [Muggleton 92]. Dans cette approche, une opération de généralisation est utilisée pour construire des règles logiques qui synthétisent les règles initiales. Toutefois, nous considérons que la lisibilité du formalisme est peu adéquate pour organiser des connaissances : la représentation n'est pas hiérarchique et il est nécessaire de bien connaître le langage utilisé (similaire à Prolog) pour pouvoir exploiter les résultats.

Conclusion

Pour conclure sur cette vue d'ensemble des travaux en classification conceptuelle, nous retiendrons l'existence de deux principales approches : l'approche classique, dont nous montrons à la section suivante qu'elle est peu adaptée à la généralisation de structures prédicatives ; l'approche de type EC, que nous détaillons à travers les différents formalismes de représentation utilisés, dans les sections 2.4, 2.5 et 2.6.

2.3 Limites de l'approche classique en classification conceptuelle pour organiser des connaissances

Les approches classiques de classification conceptuelle ont conduit à la conception de systèmes incrémentaux et de systèmes non incrémentaux. Comme le montre [Bournaud 96], ces systèmes s'avèrent en général inadaptés à la construction de classification pour l'organisation des connaissances car ils sont conçus pour des tâches de prédiction. Nous présentons dans cette section le système incrémental COBWEB [Fisher 87] et quelques uns de ses descendants, ainsi que le système non incrémental KBG[15], et montrons les limites de ces approches pour l'organisation de connaissances.

COBWEB et ses successeurs

COBWEB [Fisher 87] est un système incrémental qui utilise la notion de *concept probabiliste* (cf. section 2.2.3). L'algorithme incorpore incrémentalement les objets dans une structure arborescente, où chaque noeud représente un concept (classe d'objets), et où les concepts sont mutuellement exclusifs. L'approche est descendante, et à chaque niveau un opérateur de modification de la hiérarchie peut être appliqué :

- *placer* l'objet dans une classe existante,
- *créer* une nouvelle classe,
- *fusionner* deux classes en une classe unique,
- *diviser* une classe en plusieurs classes.

La stratégie de parcours choisie est celle du gradient (*hill-climbing*). Les opérations duales de fusion et de division permettent de simuler un retour arrière (*backtracking*). A chaque niveau, les quatre opérateurs sont testés et évalués selon le critère de *category utility*, et celui ayant le meilleur score est appliqué. L'algorithme s'arrête lorsqu'une nouvelle feuille (noeud terminal) est créée.

Le critère de *category utility* est une mesure heuristique originalement développée pour prédire le *niveau de référence* (*basic level*) des classification hiérarchiques humaines [Fisher 87, page 145]). Le niveau de référence est une notion utilisée en psychologie cognitive, qui désigne les catégories d'une hiérarchie qui sont le plus rapidement verbalisées lorsqu'elles sont présentées sous forme d'images à des sujets humains. Par exemple, la catégorie *oiseau* fait partie du niveau de référence, relativement à la catégorie plus générale *animal*, et à la

15. Nous renvoyons à la thèse de Ketterlin [Ketterlin 95] pour une description plus approfondie des principaux systèmes de l' approche classique de la classification conceptuelle.

catégorie *canari*, plus spécifique. Le critère se traduit en termes probabilistes de la manière suivante, C_k étant une classe et $A_i = V_{ij}$ un couple attribut/valeur :

- il faut maximiser la typicité des attributs d'un concept[16] exprimée par : $P(A_i = V_{ij} \mid C_k)$, probabilité que l'attribut A_i prenne la valeur V_{ij} sachant que l'objet est une instance de C_k,
- il faut maximiser la prédictivité des attributs d'un concept[17] exprimée par : $P(C_k \mid A_i = V_{ij})$ probabilité que l'objet soit une instance de C_k, sachant que A_i vaut V_{ij}.

La mesure de la qualité d'une partition d'un ensemble de classes est donnée par une combinaison de ces deux valeurs. L'algorithme correspondant est donné ci-après (algorithme 1).

Algorithme 1 COBWEB : un algorithme incrémental de classification conceptuelle

COBWEB(O, C)
- O est le nouvel objet à intégrer
- C est la classe courante, initialement la racine de la hiérarchie

intégrer O dans la classe C

mettre à jour les distributions de probabilités associées à C

si C possède des sous-classes **alors**

selon l'opérateur qui maximise CU, **faire** une action **parmi**

(1) création d'une nouvelle sous-classe de C restreinte à l'objet O

(2) placement de l'objet O dans une sous-classe C_i de C puis COBWEB(O, C_i)

(3) fusion de deux sous-classes en une classe F puis COBWEB(O, F)

(4) scission d'une sous-classe en j sous-classes puis COBWEB(O, C)

fin si

L'auteur de COBWEB, Fisher, évalue son algorithme selon quatre directions [Fisher 87] : une discussion informelle sur la qualité des hiérarchies obtenues, qui met en avant certains avantages d'une représentation probabiliste des concepts ; l'utilité de la classification pour la prédiction de la classe de nouveaux objets, évaluée quantitativement ; le coût de l'incorporation d'un nouvel objet ; le nombre d'objets nécessaires avant de converger vers une hiérarchie stable. Ceci met en avant deux points essentiels : la difficulté d'évaluer la capacité d'organisation d'une hiérarchie et le fait que l'objectif essentiel du système est de maximiser la capacité de prédiction (outre la maîtrise de la complexité algorithmique du processus, qui concerne tout concepteur).

De ce fait, COBWEB n'est pas très adapté à notre objectif d'organisation des connaissances. Il est de plus limité par d'autres aspects : premièrement la hiérarchie obtenue est dépendante de l'ordre d'introduction des objets à classer ; deuxièmement il est nécessaire de fournir une distribution de probabilité aux objets, ce qui peut être fastidieux. Enfin les concepts sont limités à des couples attribut-valeur et ne permettent pas de décrire des objets structurés.

D'autres systèmes ont été conçus par la suite sur le modèle de COBWEB. Le système ADECLU [Decaestecker 93] propose une mesure d'évaluation des opérateurs différente de la

16. Ce qui revient à maximiser la similarité des attributs au sein d'une classe.

17. Ce qui revient à minimiser la similarité entre attributs de différentes classes.

category utility, basée sur la notion de contraste entre concepts ayant le même père, mais ne permet pas de résoudre les problèmes cités ci-dessus. Le système LABYRINTH [Thompson 91] étend le langage des concepts à des objets structurés. Ketterlin en fait une étude critique dans sa thèse et propose également une extension du langage, avec la possibilité de définir trois nouveaux types d'attributs [Ketterlin 95] :

- des attributs *structurés*, qui permettent de décrire les objets composés d'autres objets, c'est-à-dire à associer à un objet o un objet o' par l'intermédiaire d'un attribut,
- des attributs *multi-valués*, qui permettent d'associer à un objet o un ensemble d'objets $\{o_1, \ldots, o_n\}$ par l'intermédiaire d'un attribut,
- des attributs *séquentiels*, qui permettent de rajouter une relation d'ordre aux attributs multi-valués.

Ces améliorations sont séduisantes, mais cette approche reste essentiellement tournée vers un objectif différent du nôtre, comme l'illustre cette citation de Ketterlin [Ketterlin 95, page 42] :

> « *Une hiérarchie de concepts a deux emplois principaux : la reconnaissance et la prédiction. La reconnaissance consiste à associer une nouvelle observation [objet] à l'un des concepts de la hiérarchie. ... Le principe [de la prédiction] est de prédire, à partir d'une description partielle d'une observation [objet], les valeurs des attributs qui ne sont pas données.* »

Principe d'un système non-incrémental, KBG

Le système KBG est basé principalement sur la définition d'une mesure de similarité sophistiquée, qui s'applique sur des objets structurés. La stratégie de classification repose alors sur le choix d'un des algorithmes statistiques utilisés en analyse de données, puisque la méthode n'impose pas un algorithme déterminé. Avec KBG, un objet est représenté par une conjonction de termes prédicatifs, composés d'un prédicat P d'arité quelconque, dont les arguments sont soit des « entités », représentées par des symboles X_i, soit des valeurs V_i ayant un type. Un terme est donc de la forme :

$$P(X_1, \ldots, X_k, V_1, \ldots, V_l)$$

La mesure de similarité entre deux objets repose sur la comparaison des termes, qui est elle-même basée sur la mise en correspondance des entités des deux objets.

Schématiquement, le processus de classification se déroule comme suit : chaque objet est initialement affecté à une classe réduite à un singleton ; puis les objets les plus similaires sont rassemblés en une ou plusieurs classes ; ensuite, les classes sont caractérisées à l'aide d'un algorithme de généralisation symbolique et les objets plus généraux obtenus par la généralisation remplacent les objets regroupés ; finalement, le processus s'arrête lorsqu'il n'y a plus qu'une seule classe.

Ces travaux sont également orientés vers la prédiction de valeurs inconnues, et l'utilisation d'algorithmes statistiques nécessite de fixer divers paramètres, dont l'influence sur les résultats est souvent importante.

Conclusion sur les méthodes classiques de classification conceptuelle

Les approches classiques se sont surtout intéressées à l'aspect inférentiel des hiérarchies construites, c'est-à-dire à la possibilité de prédire des caractéristiques relatives à de nouveaux objets. Elles proposent donc des algorithmes qui ne sont pas forcément adaptés à l'organisation des connaissances : d'une part la structure des concepts obtenus ne reflète pas forcément la structure sous jacentes aux classes du domaine ; d'autre part, l'évaluation ne s'intéresse pas vraiment à cette capacité structurante. Plus précisément, les approches classiques ne permettent pas de considérer la nature spécifique des structures prédicatives. Les classes construites ne peuvent donc pas tenir compte de certaines contraintes, comme la qualité des généralisations obtenues. L'aspect qualitatif est négligé au profit de la prédictivité.
De plus, les hiérarchies obtenues sont généralement fixes, c'est-à-dire qu'elles ne permettent pas de prendre en compte les points de vue possible sur une ensemble de classes. Elles fixent une fois pour toute la structure, sur des critères certes sophistiqués, et intuitivement séduisants, mais dont la maîtrise échappe à leur concepteur dès lors que les algorithmes sont appliqués sur de nombreux objets.

Enfin, très peu d'algorithmes prennent en compte des connaissances sur le domaine. Dans la perspective de la prédiction, cela peut se justifier par le fait que l'apprentissage consiste justement à faire émerger ces connaissances et non à les fournir auparavant. Mais cet argument n'est pas valide dans la perspective de l'organisation de connaissances.

Dans les sections suivantes, nous nous intéressons à l'approche de type « Espace de Connaissances » qui propose des solutions plus satisfaisantes pour l'organisation de connaissances.

2.4 Classification conceptuelle avec les treillis de concepts : construire toutes les classes possibles

Les problèmes posés par les approches classiques - dépendances à divers paramètres, notamment l'ordonnancement, structure hiérarchique stricte - ont conduit certains chercheurs à s'intéresser aux treillis de concepts pour la classification conceptuelle [Godin 95]. Les travaux sur la classification avec les treillis de concepts ont permis l'avénement d'une nouvelle approche, qui consiste à générer toutes les classes possibles pour éviter l'utilisation de critères de sélection subjectifs. Nous présentons dans un premier temps le formalisme des *treillis de concepts*[18], initialement défini par Wille [Wille 84]. Nous présentons ensuite deux approches utilisant les treillis de concepts : GALOIS [Carpineto 96], un système de classification conceptuelle incrémental, puis les travaux de Simon et Napoli [Simon 98] sur une approche objet pour la fouille de données utilisant les treillis de concepts pour représenter des *points de vue* sur les données.

2.4.1 Les treillis de concepts

Un treillis de concepts est défini comme un *contexte* $\mathcal{C} = (O, D, I)$, où O est un ensemble d'objets, D un ensemble de descripteurs, $I \subseteq O \times D$ une relation binaire entre O et

18. Aussi appelé *treillis de Galois*.

D. Intuitivement, chaque objet se voit associer un certain nombre de descripteurs, qui permettent de le décrire. On note oId lorsque l'objet o possède le descripteur d.

Étant donné un contexte $\mathcal{C}$, un concept est une paire (X, Y) où $X \subseteq O$ et $Y \subseteq D$ avec :

$$X = \{o \in O \mid (\forall d \in Y) oId\}$$
$$Y = \{d \in D \mid (\forall o \in X) oId\}$$

X est l'extension du concept, Y est l'intension du concept. Seules un certain nombre de paires (X, Y) représentent des concepts admissibles, puisqu'il doit y avoir une correspondance exacte entre les objets et les descripteurs. L'ensemble de tous les concepts est noté $C(O, D, I)$. On définit alors une relation de subsomption $\sqsubseteq$ sur les concepts par :

$$(X_1, Y_1) \sqsubseteq (X_2, Y_2) \longleftrightarrow X_1 \subseteq X_2$$
ou de manière équivalente $(X_1, Y_1) \sqsubseteq (X_2, Y_2) \longleftrightarrow Y_1 \supseteq Y_2$

Le théorème fondamental des treillis de concepts permet d'affirmer que $(C(O, D, I), \subseteq)$ est un treillis complet, dont l'intension de la borne inférieure d'un ensemble de concept est donnée par l'intersection des intensions des concepts. De manière duale, l'extension de la borne supérieure d'un ensemble de concepts est donnée par l'union des extensions des concepts. Ce théorème permet par la suite de ne considérer que les intensions des concepts dans les algorithmes de construction du treillis. Ce résultat a permis de mettre en oeuvre des systèmes basés sur les treillis, comme le système GALOIS.

2.4.2 Le système GALOIS : une approche incrémentale

Carpineto et Romano proposent une approche incrémentale de construction de treillis pour la classification conceptuelle, avec le système GALOIS [Carpineto 96]. L'objectif est de construire, avec l'aide d'un formalisme restreint, toutes les classes possibles sur les objets à traiter. Il est similaire à celui de Mitchell [Mitchell 82], qui proposait de construire un *espace des versions* (*version space*), consistant à générer l'ensemble de définitions de concepts possibles pour un processus d'apprentissage supervisé.

GALOIS permet de prendre en compte des connaissances du domaine, en étendant la définition de base des treillis de concepts. Pour cela, on définit un ensemble de descripteurs D^*, sur-ensemble de D, et une relation d'ordre $\subseteq_{D^*}$ sur D^*. La hiérarchie obtenue sur D^* est quelconque (non stricte). La définition de la relation de subsomption entre concepts en utilisant leur intension doit être modifiée comme suit :

$$(X_1, Y_1) \sqsubseteq (X_2, Y_2) \longleftrightarrow \forall d_2 \in Y_2, \exists d_1 \in Y_1, d_1 \subseteq_{D^*} d_2 \supseteq X_2$$

De même, le théorème fondamental est modifié, puisque l'intersection des intensions n'est plus exprimable en terme d'intersection d'ensemble : l'intersection des intensions de deux concepts (X_1, Y_1) et (X_2, Y_2) est obtenue en trouvant pour chaque paire (d_1, d_2), $d_1 \in Y_1, D_2 \in Y_2$, les descripteurs les plus spécifiques de D^* qui sont plus généraux que d_1 et d_2, puis en ne retenant que les élément les plus spécifiques parmi ceux obtenus.

L'algorithme de construction incrémental du treillis repose sur deux principes : d'une part, lorsqu'un objet est ajouté, il n'est pas nécessaire de considérer toutes les combinaisons possibles d'objets, mais seulement les concepts existants du treillis ; d'autre part, lors de l'introduction d'un nouvel objet, les concepts sont examinés en tirant partie de la relation d'ordre, ce qui permet réduire la complexité du processus. L'algorithme donné (algorithme 1) est la procédure d'introduction d'un objet dans le treillis.

Algorithme 2 Algorithme d'insertion d'un nouvel objet dans le système GALOIS

INSERTION-TREILLIS(T, o)

- T est un treillis
- o est le nouvel objet à insérer

1: NT $\leftarrow$ T ;; *NT est le nouveau treillis*
2: **pour tout** concept c de T faire **faire**
3: $c' \leftarrow c \cap o$
4: **si** non((c' vide) $\vee$ ($c' = c$) $\vee$ ($\exists parent(c) \mid parent(c) = c'$) $\vee$ ($\exists parent(c) \mid parent(c) \supset c'$)) **alors**
5: créer un nouveau noeud n dont l'intension est c'
6: NT $\leftarrow$ RELIER (n, NT)
7: **fin si**
8: **fin pour**
9: retourner NT

RELIER(n, T)

- n est un noeud
- T est un treillis

1: trouver l'ensemble CPG des concepts le plus généraux qui sont plus spécifiques que le noeud, et l'ensemble CPS des concepts les plus spécifiques qui sont plus généraux que le noeud,
2: éliminer les liens entre les éléments de CPG et CPS
3: ajouter les liens entre le noeud et chaque élément de CPG et de CPS

La complexité du processus de construction du treillis croît de manière linéaire ou quadratique en temps, en fonction du nombre de concepts, et selon divers paramètres. En pratique, le système semble atteindre sa limite autour de la prise en compte de quelques milliers d'objets.
Le système GALOIS a été appliqué à la prédiction de la classe de nouveaux objets et à la découverte de classes sur des ensemble de données couramment utilisées en apprentissage, en obtenant de bons résultats [Carpineto 93]. Toutefois le principal objectif du système est de pouvoir organiser une base de documents afin de fournir une aide à l'activité de *recherche d'information* [Carpineto 96]. Les objets du domaine représentent ainsi des documents, tandis que les descripteurs sont des mots-clés décrivant les documents. Il est alors possible de tirer partie de la visualisation graphique du treillis pour fournir un système efficace d'aide à la recherche d'informations. Nous donnons ci-après un exemple avec six documents (numérotés de 1 à 6) catalogués à l'aide de huit mots-clés : le tableau 2.1 donne les mots-clés associés à chaque document, et la figure 2.5 montre le treillis de concepts correspondant.
L'approche adoptée pour le système GALOIS, consistant à générer toutes les classes,

	1	2	3	4	5	6
intelligence artificielle	×	×	×	×	×	×
système expert	×	×	×	×	×	
recherche d'information	×					×
cataloguer		×			×	×
indexation			×			
sciences de l'information				×		
système de recherche d'information					×	
système à base de connaissances						×

TAB. 2.1 – *Une base bibliographique représentée par une matrice document/mot-clé d'après [Carpineto 96]*

s'avère pertinente pour l'organisation d'un ensemble de mots-clés : la hiérarchie obtenue n'est pas trop difficile à interpréter. Cependant, dans le cas de structures prédicatives, qui sont des unités plus sophistiquées que les mots-clés, cette solution peut conduire à une structure trop complexe, à la fois par sa taille et par l'enchevêtrement des liens entre les classes. L'interprétation devient difficile, et demande beaucoup de temps à l'utilisateur, qui perd le bénéfice d'un outil d'analyse automatique du contenu.

2.4.3 Travaux de Simon et Napoli : l'apport de la représentation par objets

Les travaux de Simon et Napoli [Simon 98] s'inscrivent dans le cadre de la *fouille de données*[19], mais leur approche a de nombreux liens avec celle du système GALOIS : un système de représentation par objets est utilisé pour représenter des connaissances expertes sur un domaine, formant une hiérarchie conceptuelle. Les données à traiter sont vues comme des instances des concepts représentés. La construction d'un treillis de Galois est alors un

19. La fouille de données est définie par Simon et Napoli comme « *l'activité qui consiste à analyser des données brutes de façon à extraire un ensemble d'unités de connaissances pouvant devenir exploitable* » [Simon 98].

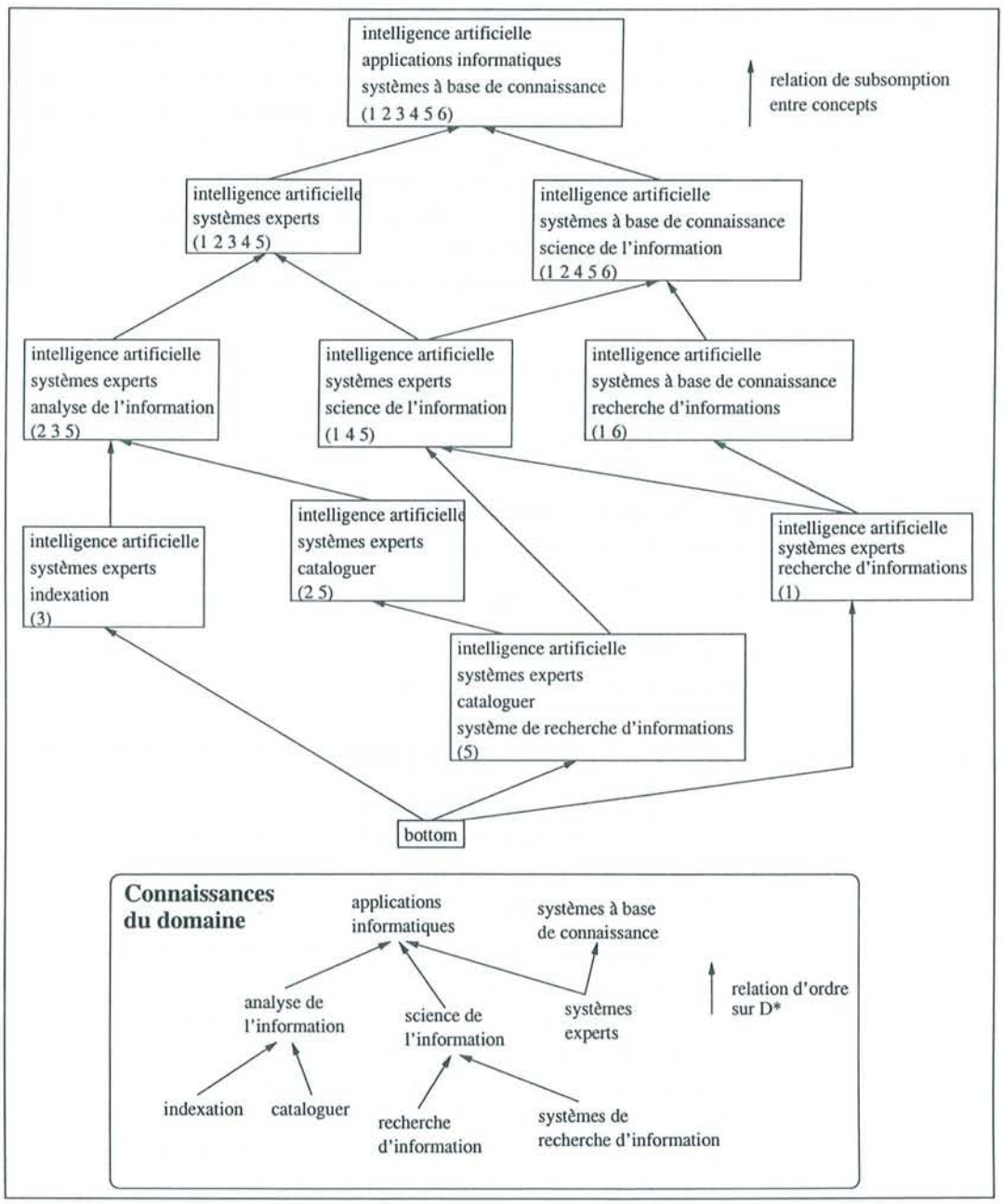

FIG. 2.5 – *Le treillis de concepts et les connaissances du domaine relatifs à six documents d'une base bibliographique d'après [Carpineto 96]*

moyen d'obtenir une nouvelle hiérarchie de classes représentant un point de vue particulier sur les données.

La représentation par objets utilise une *hiérarchie conceptuelle* telle que nous l'avons définie. L'extension d'un concept, notée $Ext(\alpha)$, est définie de manière classique par l'ensemble des objets o du domaine de référence $\mathcal{D}$ qui sont des instances de ce concept. L'intension d'un concept est décrite par une ensemble de propriétés, correspondant à des attributs prenant des valeurs dans un co-domaine. Trois types d'attributs sont possibles : un attribut symbolique *mono-valué* correspond à une valeur nominale unique ; un attribut symbolique *multi-valué* correspond à une conjonction de valeurs nominales ; enfin une *relation* met en correspondance une classe β (qui représente alors le co-domaine de la relation) avec la classe α possédant l'attribut. Une contrainte de valeurs obligatoires peut être définie pour l'attribut d'une classe. L'intension d'une classe α s'exprime alors par l'ensemble des conditions suivantes, nécessaires et suffisantes pour qu'un individu o soit une instance de la classe α :

$Att(x)$ représente l'ensemble des attributs de x (objet ou concept)
$a(o)$ désigne l'ensemble des valeurs de l'attribut a pour l'objet o
$Co - dom(\alpha, a)$ désigne l'ensemble des valeurs admissibles pour l'attribut a de la classe α (le co-domaine)
$Valeur(\alpha, a)$ désigne les valeurs obligatoires pour l'attribut a de la classe α
$Att(o) \supseteq Att(\alpha)\ et\ \forall a \in Att(\alpha) : a(o) \subseteq Co - dom(\alpha, a), Valeurs(\alpha, a) \subseteq a(o)$

La relation de subsomption est définie de manière intensionnelle (et notée $\succeq$)[20] par :

$Int(\alpha) \succeq Int(\beta) \longleftrightarrow$
$Att(\alpha) \supseteq Att(\beta) et \forall a \in Att(\beta) : Co{-}dom(\alpha, a) \subseteq Co{-}dom(\beta, a), Valeurs(\beta, a) \supseteq Valeurs(\alpha, a)$

L'ensemble de toutes les intensions qu'il est possible de construire à partir d'un ensemble $\mathcal{A}$ d'attributs, ainsi que l'ensemble de toutes les extensions qu'il est possible de construire à partir de l'ensemble $\mathcal{D}$ d'objets, forment chacun un treillis sur leurs relations respectives, $\succeq$ et $\supseteq$. Cette représentation permet d'exprimer sous forme hiérarchique des objets structurés, dépassant le cadre attribut-valeur proposé par le système GALOIS.

A partir de cette représentation, les auteurs définissent un *point de vue* comme une hiérarchie $\mathcal{H}(\mathcal{D}', \mathcal{A}')$, qui consiste en un treillis de Galois où $\mathcal{A}'$ est un sous-ensemble d'attributs de $\mathcal{A}$, et $\mathcal{A}'$ est un sous-ensemble d'objets de $\mathcal{D}'$. Chaque classe α de la hiérarchie $\mathcal{H}(\mathcal{D}', \mathcal{A}')$ est définie par le couple $(Ext(\alpha), Int(\alpha))$. En faisant varier les ensembles $\mathcal{A}'$ et $\mathcal{D}'$, il est possible d'obtenir des points de vue différents sur les données.
Pour construire les treillis de Galois, Simon et Napoli s'appuient sur l'algorithme incrémental proposé dans GALOIS, qu'ils étendent aux attributs multi-valués et aux relations. Ils utilisent également l'organisation des connaissances du domaine dans la construction des treillis, afin d'assurer que toutes les propriétés communes aux objets soient mises en évidences au sein des concepts. Un autre aspect important est la possibilité de stocker et de réutiliser les connaissances extraites, puisque connaissances du domaine et point de vue sont exprimés dans le même formalisme.
L'objectif final du système est d'extraire des règles à partir des points de vue, en s'appuyant sur des raisonnement du type : si une classe α possède en propre (sans les hériter)

20. Nous gardons la notation $\sqsupseteq$ pour la relation de subsomption entre concepts, i.e. $\alpha \sqsupseteq \beta$.

les propriétés $p1$ et p_2, et hérite la propriété p_3, alors on peut extraire, entre autres, les règles $p_1 \rightarrow p_3$, $p_2 \rightarrow p_3$ et $p_1 \wedge p_2 \rightarrow p_3$.

Ce travail a été appliqué au domaine médical, notamment sur la fouille de données épidémiologiques sur des cancers. L'objectif est de trouver des règles qui mettent en évidence des faits, par exemple, que « l'amiante est un facteur de risque des cancers c_a et c_b ». L'aspect le plus intéressant est cependant la possibilité d'organiser des connaissances exprimées en termes d'objets structurés et de calculer des points de vues. Les auteurs ont d'ailleurs choisi de présenté graphiquement la hiérarchie à l'analyste plutôt que de lui fournir une liste de règles, difficile à appréhender.

Cette approche favorise l'expressivité du formalisme au détriment d'une complexité calculatoire que l'on peut juger élevée (elle n'est pas donnée par les auteurs). Dans notre perspective d'analyse de l'information, l'exploitation de la hiérarchie implique trop de manipulations de la part de l'utilisateur : celui-ci doit choisir des points de vue et explorer différentes structures. Nous devons proposer un processus qui demande moins de travail d'exploration à l'utilisateur, car l'objectif est de l'assister dans son analyse d'un corpus de textes sans lui imposer une charge de travail annexe.

2.5 Classification conceptuelle avec les graphes conceptuels : un formalisme plus puissant

L'utilisation des graphes conceptuels pour la classification est motivée par les mêmes raisons que celles invoquées pour les treillis de concepts : l'insuffisance des approches classiques [Godin 95]. Le formalisme des graphes conceptuels [Sowa 84] présente l'avantage d'être plus expressif que les treillis de concepts. Nous verrons que cet avantage est également un inconvénient en ce qui concerne la complexité des calculs. Après une présentation du formalisme (section 2.5.1), nous nous intéressons aux deux principales approches qui ont été mises en oeuvre pour la classification conceptuelle : la Méthode de Structuration par Généralisation (MSG) de Mineau (section 2.5.2) et le système COING, une extension de la MSG proposée par Bournaud (section 2.5.3).

2.5.1 Les graphes conceptuels

Le formalisme des *graphes conceptuels* a été initialement proposé par Sowa [Sowa 84], puis a connu de nombreuses extensions par la suite [Mugnier 96]. Ce modèle s'inspire à la fois des réseaux sémantiques et de la logique, avec l'objectif d'être un système logique (et graphique) pour la représentation de la sémantique du langage naturel [Sowa 91a].

Un domaine de connaissance est représenté à l'aide d'un ensemble de *graphes conceptuels*, composés de *sommets concepts*, représentant des classes ou des individus, et de *sommets relations*, représentants des liens entre sommets concepts. Un graphe est fini, connexe et biparti (il possède deux types de sommets).

Les principales propriétés représentationnelles des graphes conceptuels sont les suivantes :

- le vocabulaire conceptuel est composé d'un ensemble de *types de concepts*, T_c, et d'un ensemble de *types de relations*, T_r. T_c et T_r sont partiellement ordonnés selon une

relation de généralité, et forment des treillis. Les types de concepts représentent des classes d'individus. Un ensemble M de marqueurs individuels représente les individus particulier du domaine représenté,

- les sommets concepts sont des paires (*type*, *référent*), composées d'un type appartenant à T_c et d'un référent, qui peut être générique (noté *) ou individuel (appartenant à M). Selon le type de référent, un sommet concept représente un *concept générique* (par exemple : [HOMME: *]) ou un *concept individuel* (par exemple: [HOMME: Jean]),
- les sommets relations sont des étiquettes nommant le type de relation, et sont liés à n sommets concepts de manière ordonnée, n représentant l'arité de la relation,
- des *signatures de relations* permettent d'exprimer des contraintes sur les types de concepts autorisés comme arguments de relations,
- un mécanisme de λ-abstraction permet l'introduction de nouveaux types de concepts, définis par genre et différence,
- un opérateur d'interprétation logique, Φ, permet la transformation de graphes conceptuels en formules bien formées de la logique des prédicats, en préservant la satisfiabilité. L'opérateur Φ fournit ainsi une sémantique basée sur la théorie des modèles au formalisme.

Les possibilités de raisonnement des graphes conceptuels reposent sur les propriétés suivantes :

- un ensemble d'opérations élémentaires sur les graphes permet de définir une relation de spécialisation (subsomption) sur les graphes, notée $\sqsubseteq$. Des opérations duales de généralisation peuvent être définies,
- l'opération de *projection*, qui est un morphisme de graphes, permet de vérifier qu'un graphe conceptuel est plus spécifique qu'un autre ($G_1 \sqsubseteq G_2$ si et seulement si il existe une projection de G_2 dans G_1),
- la structure induite sur les graphes par la relation de spécialisation est un treillis, appelé la *hiérarchie de généralisation.*

Un exemple

Avant de donner une présentation plus formelle des graphes conceptuels, nous présentons un exemple simple de base de connaissances (voir figure 2.6).

Les éléments maximum et minimum des treillis de types sont notés respectivement $\top$ et $\bot$. Le formalisme possède deux notations, une graphique et une textuelle. Selon le cas, un sommet concept est noté :

HOMME: Jean ou [HOMME: Jean]

et un sommet relation est noté (pour une relation binaire) :

→(OBJET)→ ou ->(OBJET)->

Sur la figure 2.6, le treillis des concepts fait apparaître quatre concepts, MANGER, LIEU, NOURRITURE et ETRE-VIVANT qui sont spécialisés en différents sous-concepts, tous plus généraux que $\bot$.

Le treillis des relations fait apparaître trois types de relation principales, OBJET, AGENT et LOCALISATION, cette dernière étant par la suite spécialisé par deux sous-relations. Les signatures des relations contraignent les arguments de celles-ci : par exemple, la relation AGENT ne peut accepter comme sommet concept source que des concepts plus spécifiques que le type de concept MANGER (ou des individus qui sont instance de MANGER).

Le *graphe* 1 peut être décrit par : « un homme dégustant une pizza près d'un jardin ». Dans les trois graphes, il n'y a que des concepts génériques[21]. Une situation particulière faisant intervenir l'individu Jean serait décrite en utilisant le sommet concept [HOMME : Jean].

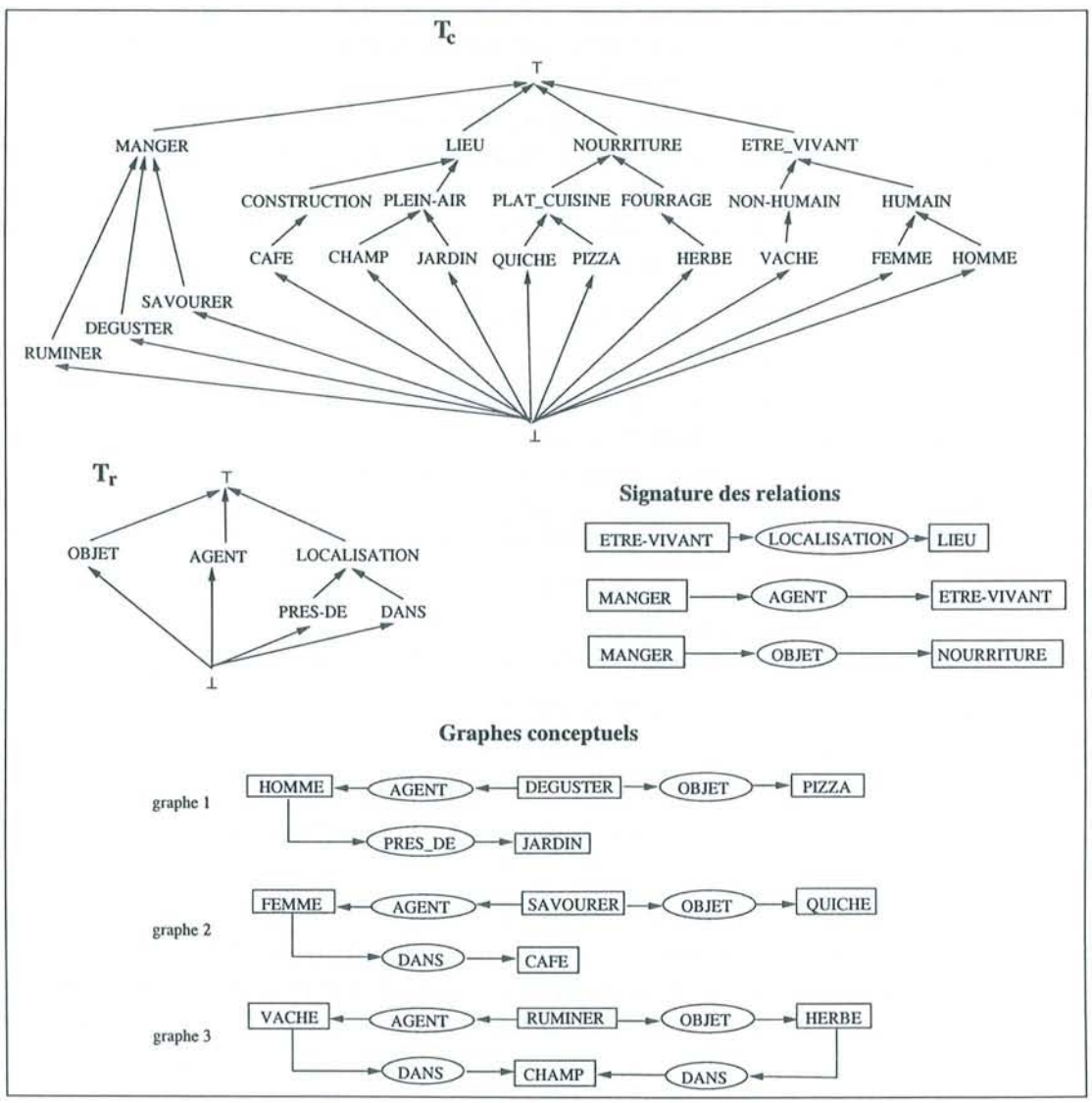

FIG. 2.6 – *Exemple de base de connaissances avec les graphes conceptuels*

21. Le référent * peut être omis dans ce cas : [HOMME] est équivalent à [HOMME : *]

Définition formelle

Nous donnons ici les éléments principaux de la définition formelle de Chein et Mugnier [Mugnier 96] pour le modèle de base, qui diffère légèrement du modèle initial de Sowa. Un graphe conceptuel est défini par rapport à un support $\mathcal{S}$. Un support est un quintuplet $\mathcal{S} = (T_c, T_r, \sigma, M, \tau)$ où :

- T_c, ensemble des types de concepts, est un ensemble partiellement ordonné, dont le plus grand élément est $\top$ (type universel) et le plus petit élément $\bot$ (type absurde),
- T_r, ensemble de types de relations, est un ensemble partitionné selon l'arité des relations, où chaque partition admet un plus grand et un plus petit élément,
- σ, qui correspond à la signature des relations, associe à tout type de relation le type maximal autorisé pour chacun de ses arguments. Pour $t_r \in T_r$, d'arité n, $\sigma(t_r)$ est un n-uplet $(T_c)^n$ et pour t_{r1}, t_{r2} d'une même partition, si $t_{r1} \sqsubseteq t_{r2}$ alors $\sigma(t_{r1}) \sqsubseteq \sigma(t_{r2})$. On note $\sigma_i(r)$ le i^{ieme} argument de $\sigma(t_r)$,
- M est l'ensemble dénombrable des marqueurs individuels. * est le marqueur générique. $M \cup \{*\}$ forme l'ensemble des marqueurs, muni d'un ordre partiel où * est plus grand que n'importe quel $m \in M$, et où $m, n \in M$ sont incomparables si $m \neq n$,
- τ est une application de M dans $T_c \setminus \{\bot\}$ qui associe un type de concept à tout marqueur (ou référent) individuel.

Étant donné un support $\mathcal{S}$, un graphe conceptuel est un multigraphe non orienté[22] biparti $\mathcal{G} = (C, R, U, etiq)$ où :

- C et R sont deux ensembles de sommets, respectivement appelés sommets concepts et sommets relations, C étant non vide,
- U est l'ensemble des arêtes ; pour tout sommet relation r, l'ensemble des arêtes adjacentes est totalement ordonné (et généralement numéroté de 1 à n, n étant l'arité de la relation). On note $G_i(r)$ le i^{ieme} voisin de r,
- *etiq* est la fonction qui associe à chaque sommet du graphe une étiquette :
  - $\forall r \in R, etiq(r) = Type(r) \in T_r$,
  - $\forall c \in C, etiq(c) = (Type(c), Referent(c)) \in T_c \times (M \cup \{*\})$
- etiq vérifie les contraintes imposées par σ et τ :
  pour tout $r \in R$, $type(G_i(r)) \sqsubseteq \sigma_i(type(r))$
  pour tout $c \in C$, si $marqueur(c) \in M$, alors $type(c) = \tau(marqueur(c))$

Les opération élémentaires de spécialisation d'un graphe G en graphe H sont les suivantes :

- la *simplification* : si G possède deux relations de même type et liées aux mêmes sommets concepts, H s'obtient en supprimant une des deux relations,
- la *restriction de relation* : H s'obtient en remplaçant le type d'une relation r de G par un type plus spécifique, à condition que les contraintes fixées par σ soient vérifiées,
- la *restriction de concept* : H s'obtient en remplaçant l'étiquette d'un sommet concept c de G par une étiquette plus spécifique (marqueur et type de concept plus spécifiques),
- le *joint interne* : H s'obtient en fusionnant deux concepts c_1 et c_2 de G ayant la même étiquette,

22. C'est-à-dire qu'il peut exister plusieurs arêtes entre un sommet concept et un sommet relation.

- la *somme* de deux graphes G_1 et G_2 : H s'obtient par somme disjointe (par juxtaposition) de G_1 et G_2.

Ces opérations définissent ensemble la relation de spécialisation présentée plus haut. La sémantique logique est donnée par l'opérateur Φ, qui transforme un graphe syntaxiquement correct en formule bien formée de la logique du premier ordre. Soit un graphe G, $\Phi(G)$ est obtenu à travers les opérations suivantes :

- associer à chaque concept générique une variable distincte x_i,
- associer à chaque marqueur individuel m la constante m,
- représenter chaque concept c de G par un prédicat unaire ayant comme nom $type(c)$et comme argument la variable ou constante associée à c,
- représenter chaque relation r, de degré n, par un prédicat n-aire ayant comme nom $type(r)$ et comme i^{eme} argument la variable ou constante associée au sommet concept lié à r par la i^{eme} arête,
- retourner la conjonction de tous les prédicats,
- fermer existentiellement les variables de la formule $(\exists x, y)$

De même, Φ est défini pour interpréter les relations de généralité définies sur les treillis de concepts et de relations T_c et T_r :

- à chaque couple $t_2 \sqsubseteq_c t_1$ de la relation de spécialisation définie sur les types de concepts, on associe la formule $\forall x,\ t_2(x) \longrightarrow t_1(x)$,
- à chaque couple $t_2 \sqsubseteq_r t_1$ de la relation de spécialisation définie sur les les types de relations, on associe la formule $\forall x_1 ... x_n,\ t_2(x_1, ..., x_n) \longrightarrow t_1(x_1, ..., x_n)$.

Chein et Mugnier ont montré qu'étant donnés deux graphes G_1 et G_2, $G_1 \sqsubseteq G_2$ si et seulement si $\Phi(G_1) \longrightarrow \Phi(G_2)$. Il faut toutefois noter que ce résultat ne s'applique qu'à une partie du formalisme (les constructions de base), constituant un sous-ensemble de la logique des prédicats. Le formalisme possède en effet de nombreuses extensions, dont l'expressivité dépasse largement celle de la logique des prédicats.

Propriétés inductives des graphes conceptuels

Les graphes conceptuels ont été beaucoup utilisés en traitement automatique du langage naturel (par exemple, [Zweigenbaum 97] [Rassinoux 94] [Barrière 96]). Cela est dû notamment à une grande expressivité, qui, selon Sowa, dépasse celle la logique des prédicats et permet de représenter les finesses inhérentes à la langue humaine [Sowa 91a]. Cette expressivité a fait le succès des GC en linguistique informatique, mais constitue un handicap pour son utilisation dans un cadre inductif.

En effet, l'approche inductive avec des objets structurés fait appel à une opération de généralisation sur ces objets. La généralisation de deux graphes nécessite tout d'abord de les apparier. Or ce problème est similaire à l'isomorphisme de sous-graphes qui est un problème NP-complet [Haussler 89]. Pour mettre en oeuvre un processus inductif, il est donc nécessaire de restreindre le langage (on parle de *biais inductif*) pour limiter la complexité de la généralisation.

Nous présentons dans la section suivante la solution proposée par Mineau, puis étendue par Gey et Bournaud, pour contourner cette difficulté.

2.5.2 La Méthode de Structuration par Généralisation (MSG) pour généraliser des objets structurés

La MSG proposée par Mineau [Mineau 90] repose sur la construction d'une hiérarchie conceptuelle appelée *Espace de Connaissances*. Les concepts (et objets) sont exprimés avec le formalisme des graphes conceptuels.
Étant donné un ensemble $\mathcal{D}$ d'objets structurés, l'EC contient l'ensemble des *plus petites généralisations communes*[23] que l'on peut calculer sur les objets de $\mathcal{D}$. Le choix des généralisations les plus spécifiques est naturel, dans la mesure où l'on cherche à trouver des concepts qui synthétisent les descriptions des objets tout en conservant un maximum d'informations. La hiérarchie construite par la MSG est considérée comme étant « objective » car elle n'élimine aucun concept sur la base d'un critère défini *a priori*. Elle rejoint en cela les approches basées sur les treillis de concepts [Carpineto 93], et s'oppose aux méthodes classiques de classification conceptuelle.

Le nombre de plus petites généralisations communes que l'on peut définir à partir de n objets est égal au nombre de partitions possibles sur ces objets. Toutefois, le nombre de concepts différents qu'il est possible de générer peut être limité par le langage. Ainsi, N attributs pouvant prendre k valeurs distinctes, le nombre de concepts est limité à $(k+1)^N$, et reste indépendant du nombre d'objets [Bournaud 96, page 73].

Comme nous l'avons vu, le formalisme des graphes conceptuels est trop expressif pour pouvoir apparier, et donc généraliser, des concepts de manière efficace. L'approche adoptée par Mineau consiste à décomposer chaque graphe en un ensemble d'*arcs indépendants*. Seules les relations binaires sont considérées, ce qui conduit à des arcs de la forme :

$$[c_1] \rightarrow (r) \rightarrow [c_2]$$

où c_1 et c_2 sont des sommets concepts et r un sommet relation. Si l'on considère une second arc' :

$$[c'_1] \rightarrow (r') \rightarrow [c'_2]$$

les seuls appariements possibles, du fait de l'orientation des arcs sont :

c_1 sur c'_1,
r sur r',
c_2 sur c'_2.

Toutefois, la décomposition des graphes conduit inévitablement à la perte d'informations dès lors que deux sommets concepts ont la même étiquette dans un graphe [Bournaud 96].

L'ensemble des arcs, obtenus par décomposition des graphes représentant les objets, sont ensuite généralisés en utilisant une règle de généralisation des types très simple : étant donné un type de concept ou de relation t_1, la généralisation de t_1 est le type générique noté « ? ». Ainsi la MSG n'exploite pas le treillis des types ou des relations généralement défini en GC. Cela revient à dire que le type générique « ? » est la racine des types de concepts et de relations, et que tous les autres types sont incomparables deux à deux. Il existe donc huit généralisations possibles pour un arc initial, chaque type pouvant être généralisé en « ? » (2^3 possibilités).
L'algorithme de généralisation est alors le suivant :

- pour chaque arc de chaque objet, appliquer le mécanisme de généralisation des types

23. Ce terme est défini page 7.

- rechercher les arcs communs à toutes les descriptions (arcs initiaux augmentés de ceux obtenus par la première étape)
- ne conserver que les arcs les plus spécifiques de l'ensemble des arcs communs

Cette approche[24] permet de généraliser des objets exprimés avec un langage très expressif, puisqu'elle prend en compte tout graphe conceptuel ayant des sommets relations binaires. Toutefois, elle ne permet pas de prendre en compte des connaissances du domaine, puisque le treillis des concepts, le treillis des relations, ainsi que la signature des relations ne sont pas exploités. La même critique a été formulé par Gey [Gey 94], qui a proposé une solution pour remédier à ce problème, et dont les travaux ont été poursuivis par Bournaud [Bournaud 96]. Nous présentons dans la section suivante le système COING proposé par Bournaud, qui en outre permet d'extraire des hiérarchies particulières à partir de la hiérarchie principale EC.

2.5.3 Le système COING : extension de la MSG pour prendre en compte des connaissances du domaine

COING[25] [Bournaud 96] est défini comme un système d'aide à la construction de hiérarchies conceptuelles pour l'organisation de connaissances. Il apporte deux améliorations principales par rapport à la MSG de Mineau : d'une part des connaissances du domaine peuvent être prises en compte ; d'autre part, il est possible d'extraire, à partir de l'espace des connaissances EC[26], des hiérarchies « allégées » en utilisant plusieurs types de critères.

La prise en compte du domaine s'effectue de manière assez naturelle en donnant une définition plus complète de la règle de généralisation des types : étant donné un type de concept ou de relation t_1, les généralisations de t_1 sont les pères de t_1 donnés par T_c (si c'est un type de concept) ou T_r (si c'est un type de relation). Il faut également fournir une règle de généralisation des référents, qui consiste à remplacer tout référent individuel par le référent générique $*$. La généralisation de tous les arcs consiste ainsi en une *phase de saturation* où les descriptions sont augmentées en utilisant les treillis de types. De plus, les signatures des relations sont prises en compte et permettent d'éliminer immédiatement tous les arcs dont la relation n'est pas conforme à la signature. En reprenant l'exemple de base de connaissances donnée dans la présentation du formalisme des graphes conceptuels, l'arc :

[DEGUSTER] → (AGENT) → [HOMME : Jean]

peut être généralisé en

[MANGER] → (AGENT) → [HOMME : Jean]

par application de la règle de généralisation de type sur le concept [DEGUSTER], mais également en

[DEGUSTER] → (AGENT) → [HOMME]

24. Nous renvoyons à [Mineau 90] ou [Mineau 95] pour le détail du processus de construction de l'espace de connaissances.

25. Pour *COnceptual clusterING*.

26. Bournaud utilise le terme *Espace de Généralisations*, mais nous conservons la désignation EC pour la cohérence de la présentation.

par application de la règle de généralisation des référents.

Pour remédier à la complexité induite par la prise en compte des treillis de types, la méthode de saturation est différente de celle utilisée dans la MSG. Les arcs sont classés en couches selon leur profondeur. La profondeur d'un arc est calculée en effectuant la somme de la profondeur de chacun des trois sommets composant l'arc, la profondeur d'un sommet étant la longueur du plus court chemin entre le type (concept ou relation) du sommet et la racine du treillis T_c ou T_r (les référents individuels n'interviennent pas dans le calcul). Ainsi l'arc

$$a = \text{[DEGUSTER]} \rightarrow \text{(AGENT)} \rightarrow \text{[HOMME]}$$

a pour profondeur $P_{arc}(a) = P_{type}(DEGUSTER) + P_{type}(AGENT) + P_{type}(HOMME)$ soit $2 + 1 + 3 = 6$. La méthode de saturation par couche débute par l'initialisation des couches, puis applique la méthode de la MSG couche par couche, en commençant par la plus profonde. Ainsi, le nombre d'arcs considéré à chaque étape est restreint, et les arcs inutiles sont éliminés sans avoir à considérer l'ensemble des descriptions.

L'algorithme de construction de l'EC s'appuie sur trois étapes distinctes :

- la saturation des descriptions (les arcs) par couche,
- la construction des noeuds de la hiérarchie, en regroupant dans un même noeud les arcs ayant la même extension (c'est-à-dire couvrant les même objets),
- la mise en place des liens de subsomption entre noeuds de la hiérarchie, en utilisant la relation d'inclusion sur les extensions.

La complexité du processus de construction utilisé par COING n'est pas plus élevée que celle de la MSG : dans le pire des cas, elle est en $\Theta(N^2)$ où N est le nombre d'objets à classifier. Le nombre de noeuds de l'espace est en $\Theta(N)$ et le nombre de liens en $\Theta(N^2)$. Cette approche est donc plus efficace que les méthodes utilisant les treillis de Galois.

L'autre apport intéressant de COING à la classification conceptuelle est la possibilité d'extraire des hiérarchies particulières à partir de l'EC. La principale motivation de cette étape d'extraction est de fournir à l'utilisateur une structure classificatoire plus simple et plus exploitable, où le nombre de concepts et de liens est réduit. Bournaud propose trois méthodes indépendantes pour y parvenir, dont deux utilisent un critère d'évaluation des noeuds C_e :

- un *élagage* itératif des noeuds les moins pertinents selon C_e, jusqu'à atteindre un seuil de pertinence ou un pourcentage de noeuds à éliminer,
- une *sélection parentale* selon C_e, qui consiste à choisir un père unique pour les noeuds ayant plusieurs pères,
- une sélection des noeuds selon un *point de vue*, qui consiste à ne conserver que les noeuds possédant une certaine caractéristique c, c étant souvent un type de relation.

Ces trois méthodes peuvent être appliquées (indépendamment) de manière automatique sur l'ensemble de la hiérarchie, aboutissant ainsi à une hiérarchie plus simple que l'EC. Il est également possible d'utiliser l'élagage et la sélection selon un point de vue de manière interactive : de cette manière, l'utilisateur peut mieux contrôler la hiérarchie finale. Bournaud propose deux principaux critères d'évaluation des noeuds C_{e1} et C_{e2}, basés sur

le nombre d'objets qui sont dans l'extension d'un noeud ($N_{ext}(n)$) et la longueur de la description d'un noeud ($N_{desc}(n)$) :

$$C_{e1} = \frac{\alpha}{N_{ext}(n)} + \beta N_{desc}(n)$$
$$C_{e2} = \alpha N_{ext}(n) + \frac{\beta}{N_{desc}(n)}$$

Pour des valeurs de α et β égales à 1, C_{e1} favorise les noeuds spécifiques, tandis que C_{e2} favorise les noeuds généraux couvrant de nombreux objets. Bournaud a appliqué le système COING à la classification de caractères chinois, ce qui a permis de montrer l'intérêt de la sélection des noeuds par point de vue : par exemple, les caractères ont pu être classés selon la relation (PRONONCIATION) et mettre en évidence différents groupes selon le type de prononciation. Une autre application, sur l'organisation de procédures comptables, a montré également l'intérêt d'une procédure d'élagage avec un critère C_e ad-hoc. Nous retenons de ces travaux qu'une hiérarchie « exhaustive » telle que l'EC est souvent trop complexe pour pouvoir être exploitée directement, et qu'il est important de proposer des hiérarchies « réduites », selon des critères pertinents pour le domaine considéré.

L'approche proposée par Bournaud avec le système COING est une solution pertinente et générale pour organiser des connaissances. Le principal inconvénient de la méthode est la perte d'informations qui découle de la décomposition des graphes : les noeuds de l'EC sont constitués par un ensemble d'arcs et non par un graphe, et il est impossible de reconstituer un graphe dans tous les cas. De plus, l'EC contient des noeuds qui correspondent à des généralisations partielles des objets initiaux, et qui contribuent à générer du « bruit » dans la structure classificatoire ainsi obtenue. A titre d'exemple, l'EC construit à partir des trois graphes de la figure 2.6, contient un noeud dont la description est l'arc :

[ETRE-VIVANT] → (LOCALISATION) → [PLEIN-AIR]

L'extension de ce noeud est composée des graphes 1 et 3, et ne correspond pas à la généralisation des deux graphes, mais à une partie seulement des deux graphes. Ce problème est dû à l'expressivité du formalisme des graphes conceptuels, et il nous semble difficile de le contourner.

2.6 Les logiques de descriptions : un meilleur compromis expressivité/complexité

Les travaux sur les logiques de descriptions n'ont pas conduit à une approche de type EC comme c'est le cas pour les treillis de concepts ou les graphes conceptuels. Pourtant, cette famille de formalismes présente des propriétés très intéressantes, car elle offre un bon compromis entre expressivité et complexité calculatoire. Après une présentation du formalisme (section 2.6.1), nous discutons de plusieurs travaux concernant les propriétés inductives des logiques de descriptions, et qui offrent des perspectives intéressantes pour la généralisation : la définition d'un opérateur de généralisation par Cohen et al. (section 2.6.2), le système KLUSTER permettant la construction inductive (section 2.6.3) et le système CANDIDE qui met en oeuvre la généralisation d'objets (section 2.6.4).

2.6.1 Les logiques de descriptions

Les *logiques de descriptions*[27] constituent une famille de formalismes issue des travaux initiaux de Brachman sur le langage KL-ONE ([Brachman 78] [Nebel 90b] [Woods 92] [Napoli 97]). Ces formalismes s'inspirent d'idées provenant de la logique des prédicats, des réseaux sémantiques et des langages de frames.

Une logique de descriptions (LD) permet de représenter un domaine de connaissances à l'aide de *concepts*, correspondant à des classes d'individus, d'*instances* de concepts, correspondant à des individus particuliers, et de *rôles* représentant des relations binaires entre individus. Les propriétés représentationnelles d'une LD sont les suivantes :

- les concepts et rôles sont exprimés à l'aide d'une description structurée, en utilisant des constructeurs dont le nombre varie selon les formalismes,
- une sémantique est associée à chaque description, par l'intermédiaire d'une fonction d'interprétation, de manière analogue à la logique des prédicats. Les manipulations syntaxiques sur les descriptions sont réalisées en accord avec cette sémantique,
- une distinction est réalisée entre le niveau terminologique ou *TBox*, relatif aux concepts, et entre le niveau assertionnel ou *ABox*, relatif aux individus,
- concepts et rôles sont organisés en hiérarchie par la *relation de subsomption*, qui les ordonne selon leur niveau de généralité. Intuitivement, un concept C subsume un concept D si l'ensemble des individus qu'il représente contient l'ensemble des individus représentés par D.

Au niveau inférentiel, une LD offre deux opérations qui constituent la base du raisonnement terminologique :

- la *classification* permet de déterminer automatiquement la position d'un concept dans la hiérarchie,
- l'*identification* permet de retrouver les concepts dont un individu est une instance.

Concepts et rôles

Les entités de base manipulées pour construire une base de connaissances en LD sont les concepts et les rôles. Chaque concept C est décrit de manière structurée à l'aide de constructeurs (*and, or, not, all, some, atmost, atleast* sont les plus courants) qui permettent d'introduire d'autres concepts, des rôles associés à C, et des *restrictions* sur ces rôles. Ces restrictions sont principalement de deux types :

- certaines portent sur le *co-domaine* du rôle, c'est-à-dire sur le concept associé par le biais du rôle,
- d'autres portent sur la *cardinalité* du rôle, c'est-à-dire sur le nombre de *valeurs élémentaires* que peut prendre un rôle, une valeur élémentaire étant soit une instance de concept, soit un type prédéfini typiquement supporté par un langage de programmation (entier, réel, caractère par exemple).

27. Désignées également par le terme *logiques terminologiques*

On distingue deux types de concepts : les concepts *primitifs* et les concepts *définis*. Les concepts primitifs sont comparables à des atomes utilisés pour construire les concepts définis. Ils n'expriment que des conditions *nécessaires* :

- si un individu i est instance du concept primitif P, alors il possède les propriétés de P.

Les concepts définis, au contraire, sont complètement caractérisés. Ils expriment des conditions *nécessaires et suffisantes* :

- si un individu i est instance du concept défini D, alors il possède les propriétés de D,
- si un individu j possède toutes les propriétés de D, alors il est reconnu comme instance de D.

De la même façon, un rôle peut être primitif ou défini. C'est l'utilisation de concepts définis qui permet de fournir une opération de classification automatique : puisque toutes ses caractéristiques sont connues, il est possible de déterminer sa position dans la hiérarchie.

Un exemple

Avant de présenter ces éléments d'une manière plus formelle, en donnant la syntaxe et la sémantique d'un langage de LD, nous donnons un exemple de description d'une petite base de connaissances en LD, où apparaissent concepts et rôles primitifs et concepts définis (figure 2.6.1).

Le concept TOP est le concept le plus général, racine de la hiérarchie. De manière identique, le rôle le plus général est *toprole*. La déclaration d'un concept primitif, d'un concept défini et d'une instance sont respectivement notés :

CONCEPT-PRIMITIF $\preceq$ *description-concept*
CONCEPT-DEFINI $\doteq$ *description-concept*
Instance :: *description-instance*

Les concepts PERSONNE, ENSEMBLE, HOMME et FEMME sont primitifs. Les différents constructeurs utilisés sont les suivants :

- **and** exprime une conjonction de concepts,
- **all** exprime la restriction du co-domaine d'un concept pour un rôle donné,
- **atleast** et **atmost** expriment des contraintes sur la cardinalité d'un rôle, respectivement le nombre minimal et maximal de valeurs élementaires que peut prendre ce rôle,
- **not** exprime la négation d'un concept (et ne s'applique qu'à un concept primitif).

On peut constater que le concept FEMME possède une description, bien qu'il soit primitif. Il est défini comme le complémentaire du concept HOMME par rapport à PERSONNE.

Les concepts représentants des équipes sont quant à eux définis : le concept EQUIPE exprime la notion d'ensemble composé d'au moins 2 membres et dont tous les membres sont des personnes ; le concept PETITE-EQUIPE spécialise la notion d'équipe en équipe n'ayant pas plus de 5 membres. Finalement, le concept EQUIPE-MODERNE spécialise la notion d'équipe en imposant un nombre de membres inférieur ou égal à 4, la présence d'au moins 1 chef, tous les chefs devant être des femmes.

PERSONNE $\preceq$ TOP
ENSEMBLE $\preceq$ TOP
HOMME $\preceq$ PERSONNE
FEMME $\preceq$ (**and** PERSONNE (**not** HOMME))
membre $\preceq$ toprole
chef $\preceq$ membre

EQUIPE $\doteq$ (and ENSEMBLE
(**all** membre PERSONNE)
(**atleast** 2 membre))

PETITE-EQUIPE $\doteq$ (**and** EQUIPE
(**atmost** 5 membre))

EQUIPE-MODERNE $\doteq$ (**and** EQUIPE
(**atmost** 4 membre))
(**atleast** 1 chef)
(**all** chef FEMME)

FIG. 2.7 – *Une petite base de connaissances en logique de descriptions d'après [Napoli 97]*

Syntaxe

Il existe de multiples langages de descriptions, dus à l'évolution des formalismes et à l'ajout successif de divers opérateurs. Nous présentons sur la figure 2.6.1 la plupart des opérateurs existants, en utilisant la syntaxe « lisp » (il existe une syntaxe « allemande », voir par exemple [Napoli 97]).

Sémantique

Le formalisme des LD associe une sémantique aux descriptions des concepts et des rôles, en utilisant la théorie des modèles de Tarski, comme pour les formules en logique des prédicats. Pour cela, on définit un ensemble d'objets $\mathcal{D}$, qui constitue le domaine de référence. Un modèle $\mathcal{M} = (\mathcal{D}, \mathcal{I})$ est constitué par un domaine $\mathcal{D}$ et une fonction d'interprétation $\mathcal{I}$, qui associe à chaque concept et chaque rôle un sens par rapport au domaine $\mathcal{D}$, ce sens correspondant à leur extension. La fonction d'interprétation permet de définir formellement, de manière extensionnelle, le sens intuitif des constructeurs. La fonction d'interprétation $\mathcal{I}$ est donnée par la figure 2.6.1 pour les constructeurs présentés.

Ce cadre théorique permet de définir formellement la relation de subsomption, qui est utilisée pour organiser les concepts et les rôles en hiérarchie. Intuitivement, un concept C subsume un concept D si pour tout modèle, l'extension de C contient l'extension de D. Plus formellement :

> Un concept C est *subsumé* par un concept D (noté $C \sqsubseteq D$) si et seulement si $\mathcal{I}[C] \subseteq \mathcal{I}[D]$ pour tout modèle $\mathcal{M} = (\mathcal{D}, \mathcal{I})$.

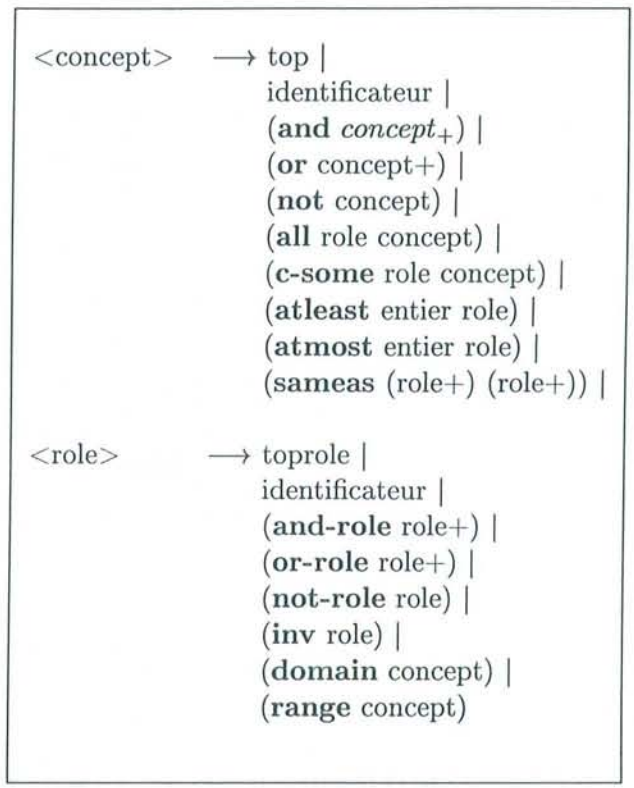

FIG. 2.8 – *Syntaxe de description d'un concept et d'un rôle en logique de descriptions*

Dans l'exemple (figure 2.6.1), le concept EQUIPE subsume à la fois les concept PETITE-EQUIPE et EQUIPE-MODERNE. De plus, PETITE-EQUIPE subsume EQUIPE-MODERNE, car toute instance du second concept est également instance du premier.

Le processus de classification

L'opération de classification consiste à trouver la position d'un nouveau concept X dans la hiérarchie $\mathcal{H}$. Elle s'opère en trois phases :

- la recherche des *subsumants les plus spécifiques* (SPS) de X,
- la recherche des *subsumés les plus généraux* (SPG) de X,
- la mise à jour des relations entre X, ses SPS et ses SPG.

Un concept C fait partie des SPS de X si C subsume X et s'il n'existe pas de concept D tel que $X \sqsubseteq D \sqsubseteq C$. De manière analogue, un concept C fait partie des SPG de X si X subsume C et s'il n'existe pas de concept D tel que $C \sqsubseteq D \sqsubseteq X$.

Le processus effectue un parcours en profondeur de la hiérarchie, en partant de la racine TOP. Lors de la première étape, si le concept courant C subsume X, il est intégré temporairement aux SPS, et sa sous-hiérarchie $\mathcal{H}_C$ est inspectée. Si un subsumant spécifique est trouvé dans $\mathcal{H}_C$, il remplace C, dans le cas contraire C est définitivement intégré. Dans le cas où C ne subsume pas X, il est écarté de la recherche, ainsi que la sous-hiérarchie correspondante.

Soient $\mathcal{C}$ un ensemble de concepts, $\mathcal{R}$ un ensemble de rôles,
Un modèle $\mathcal{M}$ est défini par $\mathcal{M} = (\mathcal{D}, \mathcal{I})$ où
$\mathcal{D}$ est le domaine de discours,
$\mathcal{I} : C \longrightarrow 2^{\mathcal{D}}$,
$\mathcal{I} : R \longrightarrow 2^{\mathcal{D}\times\mathcal{D}}$,
$\mathcal{I}$ vérifie les conditions suivantes :

Concepts

$\mathcal{I}[top] = \mathcal{D}$
$\mathcal{I}[(\textbf{and } c_1 \ldots c_n)] = \bigcap_{i=1}^{n} \mathcal{I}[c_i]$
$\mathcal{I}[(\textbf{or } c_1 \ldots c_n)] = \bigcup_{i=1}^{n} \mathcal{I}[c_i]$
$\mathcal{I}[(\textbf{not } c)] = \mathcal{D} \setminus \mathcal{I}[c]$
$\mathcal{I}[(\textbf{all } r\ c)] = \{x \in \mathcal{D} \mid \forall y(< x, y > \in \mathcal{I}[r] \rightarrow y \in \mathcal{I}[c])\}$
$\mathcal{I}[(\textbf{c-some } r\ c)] = \{x \in \mathcal{D} \mid \exists y(< x, y > \in \mathcal{I}[r] \wedge y \in \mathcal{I}[c])\}$
$\mathcal{I}[(\textbf{atleast } n\ r)] = \{x \in \mathcal{D} \mid \exists\ n\ \textit{distincts}\ y < x, y > \in \mathcal{I}[r]\}$
$\mathcal{I}[(\textbf{atmost } n\ r)] = \{x \in \mathcal{D} \mid \neg(\exists\ n+1\ \textit{distincts}\ y < x, y > \in \mathcal{I}[r])\}$
$\mathcal{I}[(\textbf{sameas } r_1\ r_2)] = \{x \in \mathcal{D} \mid \forall y < x, y > \in \mathcal{I}[r_1] \leftrightarrow < x, y > \in \mathcal{I}[r_2]\}$

Rôles

$\mathcal{I}[toprole] = \mathcal{D}^2$
$\mathcal{I}[(\textbf{and-role } r_1 \ldots r_n)] = \bigcap_{i=1}^{n} \mathcal{I}[r_i]$
$\mathcal{I}[(\textbf{or-role } r_1 \ldots r_n)] = \bigcup_{i=1}^{n} \mathcal{I}[r_i]$
$\mathcal{I}[(\textbf{not-role } c)] = \mathcal{D}^2 \setminus \mathcal{I}[r]$
$\mathcal{I}[(\textbf{domain } c] = \{< x, y > \in \mathcal{D}^2 \mid x \in \mathcal{I}[c]\}$
$\mathcal{I}[(\textbf{range } c] = \{< x, y > \in \mathcal{D}^2 \mid y \in \mathcal{I}[c]\}$
$\mathcal{I}[(\textbf{inv } r)] = \{< x, y > \in \mathcal{D}^2 \mid < y, x > \in \mathcal{I}[r]\}$

Les symboles c ou c_i, r ou r_i, et n désignent respectivement des concepts, des rôles et des entiers naturels positifs

FIG. 2.9 – *Sémantique théorie des modèles en logique de descriptions*

Les différentes logiques de descriptions

De nombreuses logiques de descriptions ont été conçues depuis le premier système de Brachman, KL-ONE. Les choix opérés par les concepteurs se sont situés autour du compromis classique entre expressivité du langage et complexité des inférences [Doyle 91]. De nombreux travaux théoriques ont permis d'atteindre des résultats très fins sur la complexité de la subsomption selon les opérateurs utilisés par un langage ([Heinsohn 92] [Woods 92] [Donini 97]). Il suffit parfois de modifier ou de rajouter un opérateur à un langage de complexité polynomiale pour basculer dans une complexité exponentielle ou un langage indécidable. Par exemple, l'opérateur **sameas** introduit une grande complexité : le langage composé uniquement de **and** , **all** et **sameas** est indécidable ! Toutefois, lorsque **sameas** est restreint à des rôles fonctionnels, c'est-à-dire n'acceptant qu'une valeur, le langage redevient décidable [Schmidt-Schauß89]

Divers systèmes ont été implémentés et sont opérationnels. Parmi les plus récents CLASSIC [Brachman 91] utilise un langage dont l'expressivité est restreinte, mais fournit des algorithmes efficaces. A l'opposé de la chaîne, LOOM [MacGregor 94] est un système qui utilise un langage très expressif, et qui reste très efficace ; toutefois, les algorithmes utilisés ne sont pas complets, et de nombreuses inférences théoriquement valides ne sont pas réalisées par LOOM.

Les propriétés inductives des logiques de descriptions

Le formalisme des LD possède des propriétés intéressantes pour un processus inductif : il permet de dépasser l'expressivité restreinte du formalisme attribut-valeur, sans atteindre la complexité de la logique du premier ordre, puisque les LD constituent une restriction de cette dernière. De fait, de nombreux auteurs mettent en avant le compromis offert entre expressivité et complexité inférentielle, qui semble être un des meilleurs que l'on puisse espérer trouver [Kietz 94] [Ventos 95]. Cette constatation n'est cependant pas applicable à toutes les LD : seules celles offrant un nombre relativement restreint d'opérateurs permettent de ne pas atteindre une complexité « critique ». Nous nous intéresserons donc aux LD de ce type. Dans la section suivante, nous présentons les travaux qui ont permis de définir l'opération de base nécessaire à un processus inductif opérant sur une LD.

2.6.2 Travaux de Cohen et al. : un opérateur de généralisation pour les LD

L'opération de généralisation, nécessaire au processus inductif en logique de descriptions, a été initialement proposée par Cohen et al. [Cohen 92]. Cette opération permet de calculer le *plus petit subsumant commun* (PPSC)[28] d'une paire de concepts, c'est-à-dire de trouver, parmi l'ensemble des descriptions qui subsument deux concepts, la description la plus spécifique. Formellement, le PPSC de deux concepts c_1 et c_2 est le concept *ppsc* tel que :

- $c_1 \sqsubseteq ppsc$, $c_2 \sqsubseteq ppsc$
- $ppsc \sqsubseteq d$ pour tout d tel que $c_1 \sqsubseteq d \wedge c_2 \sqsubseteq d$

28. En anglais, *least common subsumption* (LCS). Cette notion correspond à celle de plus petite généralisation commune que nous avons défini page 7.

Nous désignons par `ppsc` l'opérateur qui permet de calculer le PPSC de deux concepts. Cohen et al. ont défini cet opérateur pour une logique de descriptions proche de CLASSIC. Ils ont montré que pour un langage possédant l'opérateur de construction `AND` (ce qui est le cas pour toute DL), le PPSC est unique. L'opérateur `ppsc` peut être facilement étendu pour calculer le PPSC d'un ensemble de concepts, en calculant deux à deux les différents PPSC.

Ces travaux ont été poursuivis par la suite par Cohen et Hirsh [Cohen 94a] [Cohen 94b]. Les résultats, essentiellement théoriques, montrent qu'une LD contenant les opérateurs `AND` et `ALL` permet de calculer le PPSC en temps polynomial, et que l'opérateur `SAME-AS` pose des problèmes de complexité. Pour ne pas retrouver les difficultés rencontrées avec les graphes conceptuels, il est donc nécessaire de se restreindre à une LD dont l'expressivité est proche de CLASSIC sans l'opérateur `SAME-AS`. Ventos et al. ont toutefois étendu ces résultats d'apprentissage à une logique de descriptions incluant des connecteurs pour exprimer des valeurs par défaut et des exceptions [Ventos 95].

Nous montrons dans les deux sections suivantes comment l'opérateur de `ppsc`, ou un opérateur similaire, a été utilisé pour généraliser un ensemble d'objets.

2.6.3 KLUSTER : construction inductive avec une logique de descriptions

Le système KLUSTER [Kietz 94] permet l'apprentissage d'une hiérarchie de concepts à partir d'un ensemble d'assertions en LD. Sa particularité est de permettre un apprentissage constructif dans la mesure où de nouveaux termes (noms de concept) sont introduits pendant le processus. La langage utilisé par KLUSTER est voisin de celui de la LD BACK, et comprend les opérateurs **and**, **all**, **atleast**, **atmost**, **androle**, **domain**, **range** et **inverse**. Il est donc plus expressif que les langages utilisés pour la définition de l'opérateur `ppsc`.

Les données à fournir à KLUSTER sont un ensemble d'assertions correspondant au niveau de la ABox. L'exemple d'application donné par les auteurs porte sur les effets secondaires des médicaments et inclut les assertions suivantes :

```
contains(aspirin, asa)
contains(adumbran, coffein)
contains(adumbran, oxazepun)
affects(asa, headeache)
sedative(adumbran)
monodrug(aspirin)
combidrug(adolorin)
```

Il y a deux types d'assertions possibles : un prédicat binaire exprime une relation (rôle) entre deux instances du domaine, et un prédicat unaire exprime l'appartenance d'une instance à un concept. Ainsi `contains` est un rôle, `aspirin` est une instance, et `monodrug` est un concept. Les auteurs précisent que ces assertions peuvent être complétées par des connaissances du domaine (un ensemble de concepts de la TBox), mais ne développent pas ce point.

Le but est d'obtenir une hiérarchie de concepts qui organise les assertions initiales, en utilisant des descriptions aussi spécifiques que possible. Pour construire cette hiérarchie,

KLUSTER utilise trois grandes étapes :

- le calcul d'une hiérarchie de concepts et de rôles primitifs à partir des instances des concepts et rôles connus, qui permet de constituer des classes de concepts mutuellement disjoints (CMD). Chaque CMD regroupe des concepts disjoints qui possédent le même père,
- le calcul, pour chaque concept de chaque CMD, d'une *généralisation la plus spécifique* (GPS) couvrant toutes les assertions relatives au concept. Si une CMD est suffisamment discriminante, c'est-à-dire si le nombre d'assertions mal classées est inférieur à un seuil donné ϵ, elle est conservée. Dans le cas contraire, de nouveaux termes (nom de concept ou de rôle) sont introduits pour compléter la description des concepts, à condition toutefois de respecter deux paramètres gérant la longueur des descriptions,
- pour chaque CMD conservée ou complétée, le recalcul de la GPS de chaque concept si nécessaire (après augmentation du langage), puis le calcul pour chaque concept, de sa *discrimination la plus générale* (DPG), obtenue à partir de la GPS en éliminant des restrictions n'affectant pas la discrimination des assertions.

La première étape génère par exemple un concept primitif regroupant (entre autres) les deux concepts initiaux `monodrug` et `combidrug`. Ce concept doit être renommé par l'utilisateur en `drug` car le système ne peut générer qu'un nom neutre (`rootconcept_1`). De même le rôle `contains` est caractérisé :

```
monodrug ≼ drug,   ext(monodrug)={adumbran, alka_seltzer, aspirin}
combidrug ≼ drug,  ext(monodrug)={adolorin, anxiolit}
contains ≼ and domain(drug) range(substance)
```

La deuxième étape génère les GPS pour chaque concept de chaque CMD. Par exemple, les GPS des deux concepts `monodrug` et `combidrug`, appartenant à la même CMD, sont les suivantes :

```
MSG(monodrug) = and drug atleast(1, contains) atmost(2, contains)
MSG(combidrug) = and drug atleast(2, contains) atmost(3, contains)
```

L'évaluation de la discrimination d'une CMD se base sur une évaluation des GPS le constituant, prenant en compte la discrimination que permet une GPS par rapport aux autres concepts, et l'importance relative des rôles et restrictions du GPS dans cette discrimination. La CMD contenant `monodrug` et `combidrug` est par exemple insuffisamment discriminante (relativement au paramètre ϵ). Deux nouvelles relations sont alors introduites pour raffiner la CMD, conduisant à de nouvelles MSG, qui sont :

```
MSG(monodrug) = and drug atleast(1, contains) atmost(2, contains)
atleast(1, contains_active) atmost(1, contains_active)
atmost(1, contains_add_on)
MSG(combidrug) = and drug atleast(2, contains) atmost(3, contains)
atleast(2, contains_active) atmost(2, contains_active)
atmost(1, contains_add_on)
```

Les nouvelles relations `contains_active` et `contains_add_on` sont plus spécifiques, et permettent une meilleure discrimination, conduisant à une évaluation positive. Finalement, les MSG sont simplifiées et conduisent aux DPG suivantes :

```
monodrug = and drug atleast(1, contains_active)
```

```
atmost(1, contains_active)
combidrug = and drug atleast(2, contains_active)
```

Le système KLUSTER utilise un langage assez expressif, ce qui permet d'apprécier les limites à ne pas dépasser pour espérer obtenir un processus inductif efficace. Les MSG peuvent être calculées en un temps polynomial selon le nombre de concepts. Par contre, le calcul des meilleurs DPG s'avère exponentiel, et l'algorithme se contente de prendre le premier trouvé. Les opérateurs sur les rôles sont également source de complexité, et les auteurs essayent de limiter au maximum leur utilisation dans l'algorithme (« *if they are really needed* »). L'introduction de nouveaux termes pendant le processus inductif est intéressante, mais oblige l'utilisateur à nommer les termes, ce qui représente une lourde tâche dès lors que les données sont nombreuses. Enfin, la sélection des concepts générés repose sur une fonction d'évaluation dépendante de paramètres, et présente le même inconvénient que les approches classiques en classification conceptuelle.

Dans la section suivante, nous présentons deux approches qui utilisent une LD moins expressive et se basent sur l'opération de ppsc, sans reposer sur une fonction d'évaluation.

2.6.4 Utilisation de l'opération de ppsc pour généraliser des objets

Beck et al. utilisent la DL CANDIDE pour la classification conceptuelle à partir d'instances, appliquée à la conception de schémas de bases de données [Beck 94]. Ils définissent une fonction INTERSECT, qui s'avère être pratiquement équivalente à l'opérateur ppsc, et l'utilisent pour généraliser des objets (instances). CANDIDE est une DL expressive, qui met en oeuvre les opérateurs **and**, **all**, **some**, **range**, **atleast**, **atmost**. Le cadre général est le suivant : initialement, on suppose qu'il existe un schéma de base de données représenté par une hiérarchie de concepts et d'objets exprimés en CANDIDE ; une nouvelle classe ou un nouvel objet est introduit ; le résultat est une hiérarchie qui incorpore le nouvel élément. Le processus de classification conceptuelle est décrit par l'algorithme 3. La première partie

Algorithme 3 L'algorithme de classification conceptuelle de Beck et al. pour mettre à jour un schéma de base de donnée

INSERTION(E)

- E est le nouvel élement (concept ou objet) à intégrer

1: **si** E est un concept **alors**
2: utiliser l'opération de classification de la DL pour placer E
3: utiliser l'opération de réalisation de la DL pour déterminer les instances du concept
4: **fin si**
5: **si** E est un objet **alors**
6: utiliser l'opération de réalisation de la DL pour déterminer les concepts dont l'objet est instance
7: utiliser la fonction INTERSECT pour identifier les objets similaires à E
8: tester si l'objet représente une exception d'un concept existant
9: utiliser la fonction EVOLVE pour faire évoluer la hiérarchie
10: **fin si**

de l'algorithme est tout simplement l'utilisation de l'opération de classification automatique de concepts disponible dans n'importe quelle logique de descriptions. L'originalité du

processus est le traitement de l'insertion d'un nouvel objet o : celui-ci est placé selon ses caractéristiques, puis comparé à des instances d'autres concepts, en utilisant la fonction INTERSECT. Pour ne pas comparer o à toutes les autres instances de la hiérarchie, un ensemble de candidats est défini ainsi :

$CANDIDATS(o) = \{i \mid \exists r_j \in RC(o)$ tel que i possède le rôle $r_j\} \cup$
$\{i \mid \exists c$ tel que i et o sont instances de $c\}$
avec $RC(o) = \{r \mid (r$ est un rôle de $o) \vee (r$ est le père d'un rôle de $o) \vee (r$ est le fils d'un rôle de $o)\}$

L'ensemble CANDIDATS impose donc que i possède au moins un rôle ou un père en commun avec o. Cet ensemble peut être très grand en pratique. Sachant que chaque instance de CANDIDATS contribue à générer une nouvelle classe de la hiérarchie, le « bruit » (classes inutiles) généré peut être très important. Beck et al. suggèrent de guider ce processus en spécifiant a priori un but sous forme d'une nouvelle classe qui devrait subsumer tous les concepts acceptables, mais ne détaillent pas cette idée. L'utilisation de la fonction INTERSECT mériterait un processus plus contraint qui permettrait de générer des concepts de meilleure qualité.

Beck et al. ont également développé des moyens pour contourner la classification rigide en terme de conditions nécessaires et suffisantes imposée par le formalisme des DL. Le test de condition d'exception permet de classer un objet o comme instance d'un concept c s'il vérifie les conditions suivantes :

- o ne satisfait pas les propriétés de c
- le résultat de la fonction INTERSECT appliquée à o et à d'autres instances conduit à un concept c' tel que :
  $extension(c) \cap extension(c') \neq \emptyset$ et
  $extension(c) \setminus extension(c') = \{i\}$.

Une telle opération est très utile, elle implique toutefois une complexité élevée, puisque beck et al. proposent d'appliquer INTERSECT sur o et sur chaque instance différente de la hiérarchie. La procédure EVOLVE permet ensuite de créer un nouveau concept pour prendre en compte l'existence d'une exception. Une autre traitement permet de prendre en compte la notion de prototype et de valeurs par défaut. Il est toutefois beaucoup moins approfondi que la proposition faite par Ventos et al.

Le travail de Coupey et Salotti [Coupey 97] montre une utilisation plus contrainte de l'opérateur `ppsc`, dans le cadre d'un système de raisonnement à partir de cas. Le `ppsc` est appliqué sur un nombre d'instances réduit, qui doivent satisfaire un concept appartenant à une base d'index. Ceci permet de limiter la complexité du processus et de garantir une *similarité minimale* entre instances comparées. Nous retenons de ces travaux la nécessité de définir des critères plus fins que ceux proposés par Beck et al.

2.7 Conclusion : choix d'une logique de descriptions pour généraliser les structures prédicatives

Nous avons étudié tout au long de ce chapitre les avantages et les limites des différentes solutions proposées par les travaux en classification conceptuelle, problème qui s'apparente le plus à notre objectif de généralisation de structures prédicatives. Nous avons distingué

deux grands types d'approches et montré que l'« approche classique » était peu satisfaisante dans la mesure où elle privilégiait les aspects prédictifs des structures hiérarchiques obtenues, au détriment de l'aspect organisation des connaissances. De plus, l'utilisation d'une fonction d'évaluation, dont le comportement avec de nombreuses données est peu prévisible, et l'absence de prise en compte de connaissances du domaine, nous ont fait choisir une approche différente.

L'approche de type EC s'avère plus satisfaisante. Différents formalismes ont été utilisés pour mettre en oeuvre des solutions adaptées à une organisation plus « objective » des connaissances, prenant en compte des connaissances du domaine : les treillis de concepts, la représentation par objets, les graphes conceptuels. Nous avons cependant remarqué que les solutions proposées, qui s'avèrent satisfaisantes et pertinentes pour les objectifs que se sont fixés leurs auteurs, ne sont pas adaptées à notre objectif de généralisation de structures prédicatives :

- le système GALOIS (section 2.4.2), qui génère toutes les classes possibles, conduirait à une structure hiérarchique très complexe avec des structures prédicatives. Carpineto et Romano ne l'utilisent en effet que pour structurer des mots-clés, qui sont des objets très simples,
- l'approche de Simon et Napoli (section 2.4.3) propose des mécanismes pour gérer la complexité structurelle d'une hiérarchie composée d'objets complexes, avec les points de vue. Mais, dans notre cas, cela requiert un investissement trop important de l'utilisateur, qui doit lui-même générer et explorer divers points de vues,
- la MSG et le système COING (sections 2.5.2 et 2.5.3) posent le problème de la décomposition des données, qui conduirait à éclater les structures prédicatives, sans assurance de pouvoir les recomposer par la suite. De plus, les méthodes d'exploration proposées pour COING posent le même problème que l'approche de Simon et Napoli.

Nous avons enfin présenté le formalisme des logiques de descriptions, et montré qu'il offre un bon compromis entre expressivité et complexité, ainsi qu'un cadre bien défini pour une opération de généralisation. Les quelques travaux autour de la classification conceptuelle avec des logiques de descriptions ont montré qu'une approche similaire à celle de l'EC est possible. C'est ce formalisme que nous avons choisi comme cadre pour la généralisation de structures prédicatives, et plus particulièrement le système CLASSIC, pour son expressivité suffisamment restreinte pour un processus inductif [Capponi 97a]. Pour clore ce chapitre, nous énumérons les avantages d'un tel choix :

1 une LD offre un cadre formel, avec une sémantique clairement définie,

2 il existe un opérateur bien défini pour la généralisation (PPSC), ne posant pas de problème de complexité calculatoire,

3 l'opération de classification automatique (déductive) permet une mise à jour efficace de la hiérarchie, et allège ainsi le travail nécessaire au processus de généralisation,

4 la représentation des structures prédicatives nécessite de travailler au niveau terminologique, ce qui est facilité par la séparation des connaissances en TBox et ABox,

5 il existe un système opérationnel et d'expressivité adéquate : CLASSIC

Dans le prochain chapitre, nous présentons plus précisément la logique de descriptions choisie, CLASSIC, et montrons comment nous utilisons ce formalisme pour la représentation des structures prédicatives.

3

Proposition de représentation des structures prédicatives en CLASSIC

Nous avons choisi le formalisme des logiques de descriptions comme cadre pour la généralisation de structures prédicatives. Parmi les LD existantes, notre choix s'est porté sur CLASSIC dont le compromis expressivité/efficacité semble le plus approprié. Dans ce chapitre, nous présentons dans un premier temps les particularités de CLASSIC (section 3.1). Puis nous nous intéressons aux propriétés des structures prédicatives que nous utilisons, et montrons comment elles sont représentées en CLASSIC, phase préalable à la définition du processus de généralisation (section 3.2). Nous donnons une définition minimale des structures prédicatives et nous focalisons peu sur l'aspect linguistique, qui est discuté de manière plus approfondie au chapitre 5.

3.1 Présentation de la logique de descriptions CLASSIC

CLASSIC est un système implémentant une logique de descriptions [Brachman 91], [Borgida 89] [Resnick 95]. Il est axé sur la simplicité du langage de description de concepts et d'objets, que nous présentons figure 3.1. En effet, il n'existe pas de constructeur **or**, **not** ou **some** pour décrire un concept. De plus, les rôles ne peuvent être que primitifs, puisqu'il n'existe pas de constructeur de rôles. En revanche, ceux-ci peuvent être organisés en hiérarchie.

L'expressivité restreinte du langage permet d'obtenir une LD efficace en terme d'inférences, qu'elles soient déductives comme le test de subsomption et la classification (étude comparative de [Baader 94]) ou inductives comme la généralisation [Cohen 94a]. L'algorithme de subsomption utilisé est incomplet [Resnick 95], toutefois les cas où la subsomption n'est pas detectée sont restreints par rapport aux autres LD dont l'efficacité est comparable [Heinsohn 92].

Nous ne rappelons pas ici la signification des principaux constructeurs présentés section 2.6. Les constructeurs spécifiques à CLASSIC sont ceux qui permettent d'introduire des objets individuels dans les descriptions : **fills** et **one-of**. CLASSIC est une DL un peu particulière car elle ne fait pas une séparation nette entre concepts et objets : ces derniers peuvent apparaître dans la définition des concepts. Ainsi, **fills** permet de donner une ou

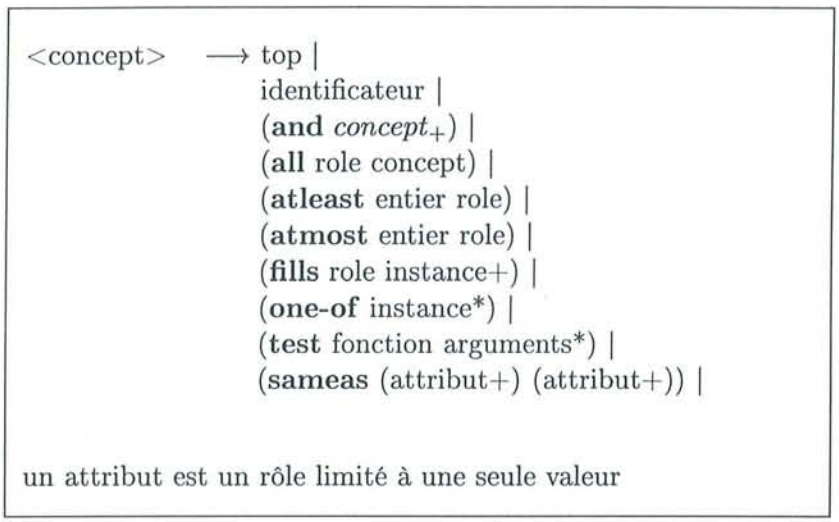

FIG. 3.1 – *Syntaxe simplifiée de description d'un concept en* CLASSIC *d'après [Resnick 95]*

plusieurs valeurs (objets individuels) à un rôle. Par exemple, pour exprimer le concept d'un vin de Bordeaux rouge, on peut utiliser l'expression

```
(and BORDEAUX (fills couleur Rouge))
```

où `BORDEAUX` est un concept représentant un vin de Bordeaux, `couleur` est un rôle et `Rouge` est un objet. Le constructeur **one-of** permet d'exprimer une disjonction de valeurs. Par exemple, pour exprimer la couleur d'un vin, on peut utiliser l'expression

```
(all couleur (one-of Rouge Rose Blanc)),
```

qui signifie que `couleur` peut prendre une ou plusieur des trois valeurs `Rouge`, `Rose` ou `Blanc`.

Une autre particularité de CLASSIC est l'opérateur **test** qui permet d'introduire un comportement procédural pour pallier les problèmes d'expressivité : une fonction est appliqué sur le concept ou l'objet, elle doit retourner la valeur VRAI ou FAUX, permettant d'exprimer une contrainte particulière sur l'objet. Ce mécanisme est toutefois limité car il ne peut être pris en compte pour la classification automatique d'un concept dans la hiérarchie. Par exemple, le concept d'entier pair peut s'exprimer par l'expression

```
(and ENTIER (test parite))
```

`ENTIER` étant un concept et `parite` une fonction Lisp retournant VRAI lorsque l'argument est un entier pair.

Comparaison entre CLASSIC et les autres formalismes de représentations

Nous avons présenté dans le chapitre précédent différents formalismes de représentation susceptibles de satisfaire notre problème de représentation des structures prédicatives. Nous comparons brièvement la LD CLASSIC avec trois formalismes : la logique du premier ordre, les représentations à objet et les graphes conceptuels.

Il existe un fort parallèle entre logique des prédicats et LD car les deux formalismes utilisent une structure sémantique similaire, la théorie des modèles de Tarski [Napoli 97]. Cependant une LD est plus restrictive car elle n'utilise pas de *variables* : les expressions de concepts et d'assertions peuvent être vues comme des cas particuliers de formule logique où ne figurent que des instances de variables. Les prédicats (au sens logique) utilisés dans une LD sont de plus limités aux prédicats unaires (concepts) et binaires (rôles). CLASSIC simplifie également l'expressivité du formalisme en omettant des connecteurs tel que la disjonction (**or**) ou la négation (**not**). Nous renvoyons à [Borgida 96] pour une comparaison détaillée de l'expressivité des deux formalismes.

Les représentations à objet utilisent les notions de *classe* et de *frame*, qui encapsulent une description sous forme d'attributs et un comportement sous forme de fonctions activées par envoi de messages. Les descriptions sont similaires à celles utilisées par une LD, et font le même usage de l'héritage des propriétés selon une structure hiérarchique. Les représentations à objet se distinguent par leur caractère nettement plus *procédural*, qui ne permet pas de proposer une sémantique formelle aux expressions du langage, mais offre plus de souplesse en terme d'expressivité [Napoli 97]. CLASSIC propose toutefois l'opérateur **test** qui permet d'introduire une composante procédurale dans une LD. La dernière version de CLASSIC permet également d'intégrer ce comportement au mécanisme de classification, mais cela relève de la programmation avancée du système. Enfin un langage à objets n'offre pas, en général, de mécanisme de classification automatique car il ne considère pas les propriétés suffisantes des concepts. Certains travaux ont toutefois défini un tel mécanisme pour la classification d'instances [Ducournau 96].

Les graphes conceptuels ont une origine commune avec le formalisme des LD, puisqu'ils proviennent tous deux des réseaux à héritage, et permettent d'organiser des concepts en hiérarchie. Les graphes conceptuels permettent toutefois une représentation beaucoup plus fine des concepts décrits, grâce à leur grande expressivité. Des phénomènes complexes inhérent au langage naturel peuvent ainsi être exprimés : quantifications élaborées, distinction subtile entre individus et concepts génériques, nombreux types de référents, représentation de propositions, ... [Sowa 91a]. Les graphes conceptuels offrent des opérations biens définies sur les graphes, qui permettent des manipulations fines sur les concepts. En contrepartie, ils n'offrent pas de classification automatique de concepts, et le modèle, très théorique, pose des problèmes de mise en oeuvre : les systèmes existants n'offrent que certaines fonctionnalités du modèle (voir par exemple [Munday 95]). Par contraste, une LD offre un langage bien défini et relativement simple. En particulier CLASSIC est un système opérationnel et simple muni d'un ensemble de fonctionnalités permettant de créer, manipuler et interroger une base de connaissances exprimée avec un langage de descriptions. Quelques travaux se sont attachés à comparer les deux formalismes de manière théorique ([Ounis 95]) ou pratique ([Nobecourt 98]).

Nous nous intéressons à présent aux spécificités des structures prédicatives que nous manipulons et à leur représentation en CLASSIC.

3.2 Méthode de représentation des structures prédicatives en CLASSIC

Nous présentons en premier lieu de façon plus détaillée la notion de *structure prédicative* que nous utilisons. Celle-ci s'inspire bien évidemment de la notion classique utilisée en

linguistique, mais en diffère sur certains points, notamment par un moindre degré de sophistication : une structure prédicative est pour nous un moyen de représenter de manière concise et structurée une unité d'information extraite des textes. Elle s'appuie sur un prédicat, qui représente une action ou un évenement du domaine considéré, et ses arguments, qui représentent des objets (au sens très général) sur lesquels s'appliquent l'action.

Les structures prédicatives

La forme générale que nous utilisons pour une structure prédicative est la suivante :

$predicat(rel_1 : arg_1, rel_2 : arg_2, \ldots, rel_n : arg_n)$
avec $arg_i = terme$
ou $arg_i = predicat'(rel'_1 : arg'_1, \ldots)$

Les élément de base composant une structure prédicative sont les *prédicats* ou *têtes prédicatives* et les arguments simples (arg_i) qui apparaissent dans les structures : nous les appelerons *termes*[29]. Le lien entre un prédicat et un argument est réalisé par une relation (rel_i). Nous notons $\mathcal{T}$ l'ensemble des termes, et $\mathcal{P}$ l'ensemble des prédicats. Les prédicats forment un sous-ensemble des termes : $\mathcal{P} \in \mathcal{T}$. L'ensemble des structures prédicatives est noté $\mathcal{SP}$.

Nous considérons qu'un prédicat $p \in \mathcal{P}$ représente une notion unique, et ne peut être considéré comme un lexème pouvant posséder plusieurs sens. Par exemple, nous nous interdisons d'utiliser le prédicat *voler* pour représenter à la fois l'action de se déplacer dans les airs et l'action de dérober un objet. Nous utiliserons deux prédicats distincts, par exemple *voler* pour le premier sens) et *dérober* pour le second. Ceci s'applique également aux autres termes.

Nous n'imposons *a priori* aucune contrainte sur le nombre de relations utilisables ou le sens donné à ces relations. Les relations doivent permettre d'exprimer des propriétés de nature différentes. Les relations thématiques classiques sont les plus appropriées [Saint-Dizier 95][30]. Nous en utiliserons une ensemble restreint avec les significations suivantes :

agent : qui est à l'origine de l'action,
objet : sur lequel porte l'action (thème),
moyen : ce qui est utilisé pour réaliser l'action,
localisation : au sens large, où se déroule l'action,
but : ce pourquoi est réalisée l'action.

Nous considérons qu'une relation ne peut être utilisée qu'une seule fois dans une structure prédicative donnée, pour des arguments de même niveau (cela ne s'applique pas si une structure prédicative est imbriquée dans une autre). Cette contrainte est souvent appliquée pour les rôles thématiques, car il est considéré que deux arguments ne peuvent jouer le même rôle sémantique [Fillmore 68]. Pour notre part, nous considérons que cette contrainte permet de simplifier la représentation, tout en conservant une expressivité suffisante.

29. Nous utilisons ici la notion de *terme* comme synonyme d'élement d'une formule ou d'une expression, sans sa connotation linguistique.

30. Les relations thématiques sont présentées de façon plus détaillées section 5.3.3.

Représentation en CLASSIC

Nous posons comme hypothèse l'existence d'un ordre partiel sur les termes, permettant de les situer les uns par rapport aux autres. En CLASSIC, nous choisissons de représenter les termes par des concepts, et la relation de subsomption permet d'exprimer l'ordre sur les termes. Les concepts représentant les termes peuvent être primitifs ou définis, selon la finesse des informations dont on dispose a priori. Dans la suite de ce mémoire, nous nous limitons à des concepts primitifs, sans propriétés, pour plusieurs raisons : premièrement, notre objectif de généralisation ne nécessite pas l'emploi de concepts définis. Deuxièmement, l'utilisation de concepts définis nécessite un travail très important de modélisation, qui n'est pas justifié pour l'utilisation que nous faisons des structures prédicatives.

Nous caractérisons une structure prédicative *sp* à l'aide d'un concept défini SP, comme la conjonction du concept PREDICAT représentant la tête prédicative et un ensemble de rôles représentants les couples relation/argument (rel_i, arg_i) de la structure prédicative[31] :

```
SP ≐ (and PREDICAT
          (all rel1 ARG1)
          (all rel2 ARG2)
          ...
          (all reln ARGn))
```

Par exemple, la structure prédicative

dosage(objet : amine, moyen : chromatographie liquide)

est représentée en CLASSIC par le concept DOSAGE-1[32] :

```
DOSAGE-1 ≐ (and DOSAGE
                (all objet AMINE)
                (all moyen CHROMATOGRAPHIE_LIQUIDE))
```

où objet et moyen sont des noms de rôles représentant les relations *objet* et *moyen*. Les trois concepts primitifs permettant de construire la structure prédicative sont DOSAGE, AMINE et CHROMATOGRAPHIE_LIQUIDE, les deux derniers étant utilisés comme restriction de rôles.

Notre représentation de structures prédicatives se situe à un niveau terminologique, par opposition au niveau assertionnel : nous les définissons en effet à l'aide de concepts, plutôt que par des objets individuels. Une alternative pourrait consister à utiliser des instances de concepts pour les définir. En reprenant l'exemple précédent, la structure prédicative dosage-1 serait définie par :

```
DOSAGE-1 :: (and DOSAGE
                 (fills objet AMINE-1)
                 (fills moyen CHROMATOGRAPHIE_LIQUIDE-1))
```

où DOSAGE est également un concept primitif, AMINE-1 et CHROMATOGRAPHIE_LIQUIDE-1 étant deux instances respectives des concepts amine et chromatographie_liquide. Cependant cette solution conduirait à l'introduction d'un nombre important d'objets individuels.

31. Le *rôle* décrit le couple (rel_i, arg_i) ; rel_i est le *nom de rôle* et arg_i la *restriction* du rôle. Lorsqu'il n'y a pas d'ambiguité rel_i est désigné simplement par le terme *rôle*.

32. Par convention, nous nommons le concept représentant une structure prédicative par le nom de sa tête prédicative suivie d'un nombre.

Plus fondamentalement, notre objectif n'est pas de nous intéresser aux particularités de chaque structure prédicative extraite d'un texte, mais de synthétiser des informations. Nous nous focalisons donc sur les concept génériques : lorsque le terme *amine* apparaît dans une structure prédicative, nous retenons le concept `AMINE` et non une instance particulière de ce concept.

Les arguments des structures prédicatives

Une question importante est soulevée par l'utilisation de structures prédicatives : le statut des arguments. En linguistique, un prédicat se voit associer une arité, qui correspond au nombre d'arguments obligatoires associés au prédicat pour former une structure prédicative. C'est-à-dire que pour un prédicat d'arité n, la structure prédicative correspondante doit posséder au minimum n arguments, éventuellement complétés par des arguments optionnels. Toutefois les structures prédicatives susceptibles d'être extraites des textes ne respectent pas forcément ces conditions : il existe de nombreux cas où les informations relatives à une structure prédicative sont éclatées à différents endroits du texte, et très difficile à recomposer. Il est donc possible d'obtenir des structures prédicatives incomplètes. Nous illustrons ces propos avec le prédicat *dosage*.

Le prédicat *dosage* peut être considéré comme un prédicat à trois arguments obligatoires : un agent, a priori humain, qui effectue l'action, un objet qui est dosé, correspondant souvent à une substance chimique, et une localisation de l'objet dosé, correspondant souvent à un produit alimentaire ou un végétal. Un quatrième argument, le moyen utilisé pour le dosage, peut-être considéré comme optionnel. Or, dans le corpus utilisé pour notre application, l'agent n'apparaît jamais dans les textes, car il est sous-entendu. De même, il est courant que l'objet ou la localisation soient précisés dans des phrases différentes de celle où apparaît le prédicat, ce qui rend très difficile une extraction complète de la structure prédicative correspondante. Nous trouvons par exemple la structure prédicative `DOSAGE-1` dont nous rappelons ici la définition en CLASSIC :

```
DOSAGE-1 ≐ (and DOSAGE
                (all objet AMINE)
                (all moyen CHROMATOGRAPHIE_LIQUIDE))
```

Pour respecter l'arité du prédicat *dosage*, il peut être pertinent de définir `DOSAGE-1` de manière plus complète en ajoutant les rôles `agent` et `localisation` avec des valeurs par défaut :

```
DOSAGE-1 ≐ (and DOSAGE
                (all objet AMINE)
                (all agent PERSONNE)
                (all localisation PRODUIT)
                (all moyen CHROMATOGRAPHIE_LIQUIDE))
```

Les concepts `PERSONNE` et `PRODUIT` correspondent aux concepts « par défaut » associés aux rôles `agent` et `localisation` pour le prédicat *dosage*[33]. Nous n'avons pas retenu cette approche, afin de proposer un processus de généralisation qui soit suffisamment général pour prendre en compte des structures prédicatives non homogènes. En effet, si le processus que nous proposons (décrit au chapitre suivant) s'applique à des structures prédicatives

33. En linguistique informatique, ce sont les *restrictions de sélection.*

n'ayant pas forcément le même nombre de relations, il s'applique *a fortiori* dans le cas particulier où les relations sont homogènes pour un même prédicat. Et dans ce cas, le résultat obtenu est potentiellement plus satisfaisant.

Représentations utilisées dans les autres travaux

La représentation adoptée considère les prédicats comme des concepts. Cette approche est suivie par de nombreux auteurs : par exemple dans le système XTRA [Jansen-Winkeln 91], la hiérarchie conceptuelle possède les deux concepts *THING* et *PREDICATE* pour classer respectivement les objets et les prédicats. Des relations thématiques (*subject, agent, destination, source*) permettent d'exprimer les liens entre les deux. Forster [Forster 94]propose également cette solution pour représenter la sémantique lexicale des verbes relatifs à des activités et des événements avec la logique de description ALAN . Pour représenter des groupes nominaux dans un réseau sémantique "à la KL-ONE", Biebow et Szulman introduisent des prédicats et des rôles thématiques comme intermédiaires : ainsi, *secrétaire bilingue* est interprété comme le concept *secrétaire sachant parler deux langues.* Le lien entre le prédicat *parler* et les deux arguments sont respectivement *agent* et *objet* [Biebow 91].
Une approche alternative consiste à considérer le prédicat comme une relation entre plusieurs objets. Cependant, elle est difficile à réaliser avec une logique de descriptions, car dans celle-ci les relations ne peuvent être que binaires. Une solution pour rendre compte d'un prédicat soit comme concept, soit comme relation a été proposée par [Franconi 94]. Il s'agit d'introduire la possibilité de réifier les relations, c'est-à-dire de créer une correspondance entre une relation et sa représentation au moyen d'un concept. Par exemple, la relation *peint* entre les concepts *artiste* et *oeuvre* peut être transformé en concept `peindre`, lié respectivement à `artiste` et `peintre` par deux relations `agent` et `sujet`.

Un exemple de représentation des structures prédicatives

Les structures prédicatives étant introduites en CLASSIC comme des concepts définis, le mécanisme déductif de classification de la LD permet une réorganisation automatique de la hiérarchie à chaque introduction d'une nouvelle structure. Par définition, chaque concept représentant une structure prédicative est classé sous le concept représentant sa tête prédicative. Les positions relatives des structures prédicatives relatives à un prédicat donné sont ensuite dépendantes de leurs caractéristiques (c'est-à-dire de leurs arguments). Ceci est illustré par la figure 3.2, montrant la hiérarchie obtenue après l'introduction de six structures prédicatives ayant la tête prédicative *dosage*, numérotées de 1 à 6.
Les relations de subsomption existant entre les différents arguments sont illustrées par la figure 3.3. La description textuelle des six concepts en CLASSIC est la suivante :

```
DOSAGE-1 ≐ (and DOSAGE
                (all objet HISTAMINE)
                (all moyen CLHP))
DOSAGE-2 ≐ (and DOSAGE
                (all objet SEROTONINE))
DOSAGE-3 ≐ (and DOSAGE
                (all objet CHOLINE)
                (all moyen CLHP))
```

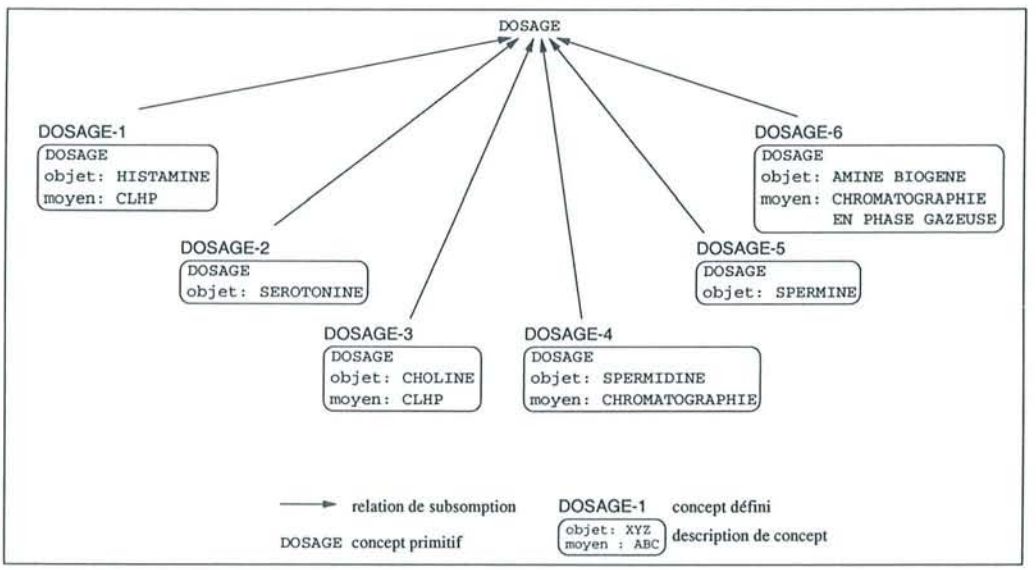

FIG. 3.2 – *Six structures prédicatives avec la tête prédicative dosage en* CLASSIC

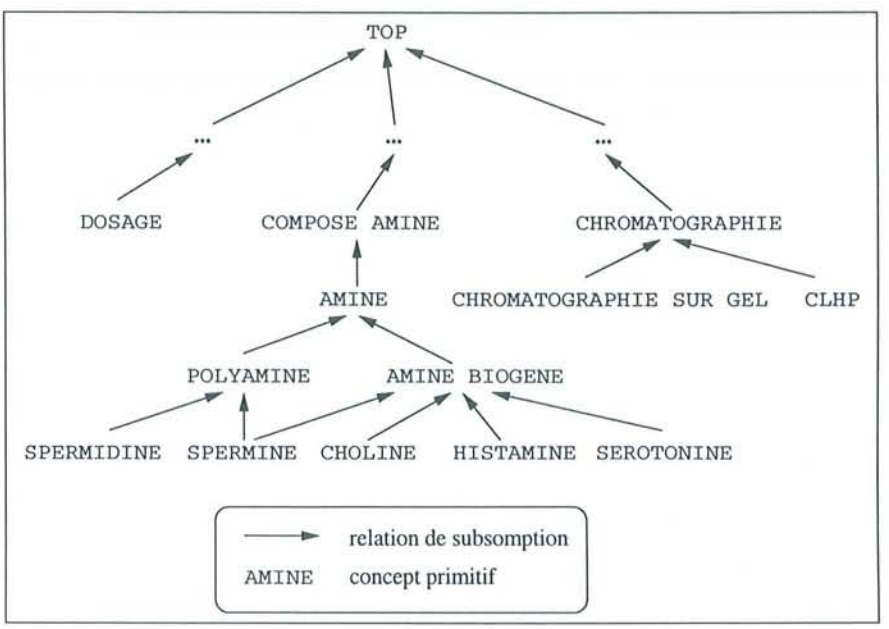

FIG. 3.3 – *Détails de la hiérarchie des concepts, avec focus sur les arguments utilisés par les structures prédicatives de dosage*

```
DOSAGE-4 ≐ (and DOSAGE
                (all objet SPERMIDINE)
                (all moyen CHROMATOGRAPHIE))
```

```
DOSAGE-5 ≐ (and DOSAGE
                (all objet SPERMINE))
DOSAGE-6 ≐ (and DOSAGE
                (all objet AMINE_BIOGENE)
                (all moyen CHROMATOGRAPHIE_EN_PHASE_GAZEUSE))
```

La figure 3.4 montre la même hiérarchie après l'introduction d'une septième structure prédicative DOSAGE-7 :

```
DOSAGE-7 ≐ (and DOSAGE
                (all objet SPERMIDINE)
                (all moyen CHROMATOGRAPHIE_SUR_GEL))
```

DOSAGE-7 se trouve ainsi placé sous le concept DOSAGE-5, celui-ci étant plus général car il possède une propriété de moins.

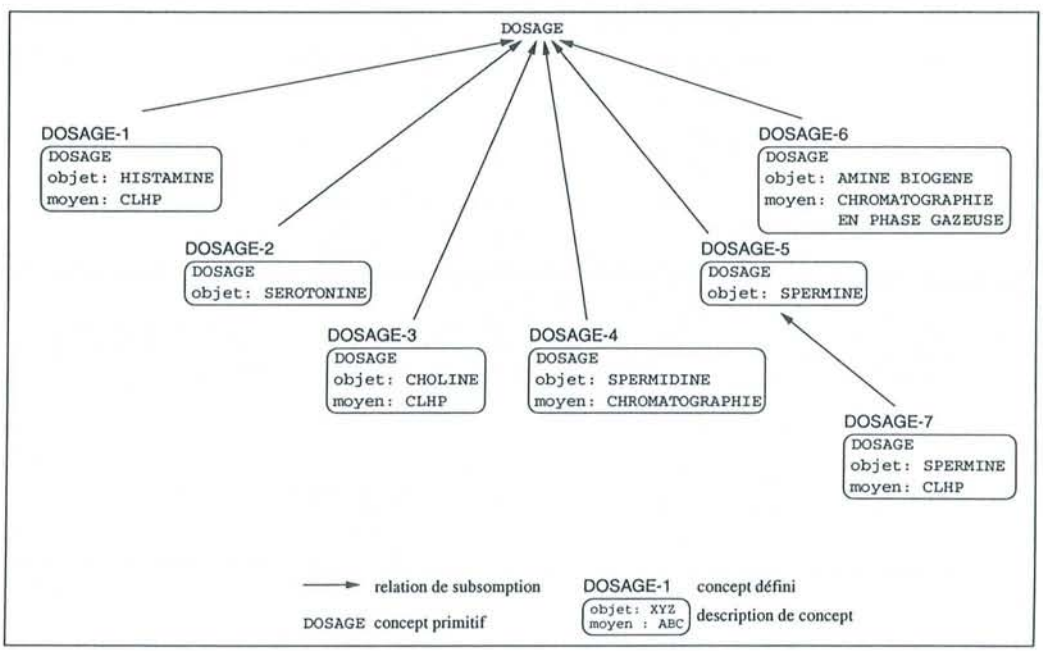

FIG. 3.4 – *Hiérarchie après l'ajout d'une septième structure prédicative en* CLASSIC

Dans la suite de ce mémoire, et particulièrement au chapitre suivant, nous serons amenés à manipuler des structures prédicatives et des concepts permettant de les représenter. Par souci de simplification, nous utiliserons souvent :

- le terme *structure prédicative* pour désigner le *concept représentant une structure prédicative*,
- le terme *prédicat* pour désigner le *concept représentant un prédicat.*

Ces simplifications seront utilisées dans le mesure où il n'y a pas d'ambiguité.

Conclusion

Dans ce chapitre, nous avons présenté la logique de description utilisée, CLASSIC, et nous l'avons brièvement comparé aux principaux formalismes utilisés pour la représentation de connaissances. Nous avons montré comment utiliser la LD CLASSIC comme cadre pour la représentation des structures prédicatives. Nous avons volontairement laissé de côté les problèmes posés par le passage de textes écrit en langage naturel aux structures conceptuelles d'une langage de représentation[34]. Nous nous sommes focalisés sur la représentation, en essayant de montrer comment sont structurées les informations représentées au moyen des structures prédicatives. Nous disposons à présent de tous les élément pour présenter notre méthode de généralisation de structures prédicatives, qui fait l'objet du chapitre suivant.

34. Ces problèmes sont discutés dans la deuxième partie, aux chapitres 5 et 6.

4

Notre méthode de généralisation de structures prédicatives

Nous proposons dans ce chapitre une méthode de généralisation de structure prédicatives basée sur la logique de descriptions CLASSIC [Capponi 98a] [Capponi 98b]. Cette méthode utilise la notion de plus petit subsumant commun (PPSC, cf. section 2.6.2) pour générer un ensemble de structures prédicatives plus générales que celles fournies initialement, et que nous appellerons *généralisations*. Dans un premier temps, nous présentons en détail l'algorithme de calcul du PPSC(section 4.1), puis nous montrons comment sont introduites dans la hiérarchie les généralisations calculées (section 4.2). Dans un deuxième temps, nous présentons le processus de généralisation proprement dit, qui consiste à appliquer l'opération de calcul du PPSC sur des ensembles de structures prédicatives « bien choisis ». Nous détaillons ainsi les heuristiques utilisées (section 4.3), puis les deux étapes principales dont est constitué l'algorithme (sections 4.4 et 4.5). Enfin nous nous intéressons à la complexité du processus de généralisation présenté (section 4.6) avant de conclure sur notre méthode de généralisation (section 4.7).

4.1 Un algorithme de calcul du PPSC

La notion de PPSC a été présentée section 2.6.2, où nous avons donné une définition pour deux concepts. Nous l'étendons à présent à un ensemble C de concepts (non réduit au singleton) :

$\forall c \in C, c \sqsubseteq ppsc$,
$ppsc \sqsubseteq d$ pour tout d tel que $\forall c \in C, c \sqsubseteq d$

L'opérateur de calcul du PPSC peut désormais s'appliquer à un ensemble de n arguments, et s'exprime facilement à partir de l'opération sur deux arguments :

$$\texttt{ppsc}(c_1, \ldots, c_n) = \texttt{ppsc}(c_n, \texttt{ppsc}(c_{n-1}, \ldots \texttt{ppsc}(c_2, c_1) \ldots))$$

L'unicité du résultat est assurée par l'existence du connecteur **and**. En effet, supposons qu'il existe deux PPSC a et b distincts pour un ensemble C, alors la conjonction (**and** a b) est plus spécifique que a et b et subsume chaque concept de C : le PPSC de C est unique et a pour valeur (**and** a b).

Un exemple de calcul de PPSC, entre les concepts DOSAGE-1 et DOSAGE-3, est illustré par la figure 4.1. Nous rappelons les relations de subsomption entre les différents concepts apparaissant dans les structures prédicatives à la figure 4.2. Le PPSC des concepts DOSAGE-1

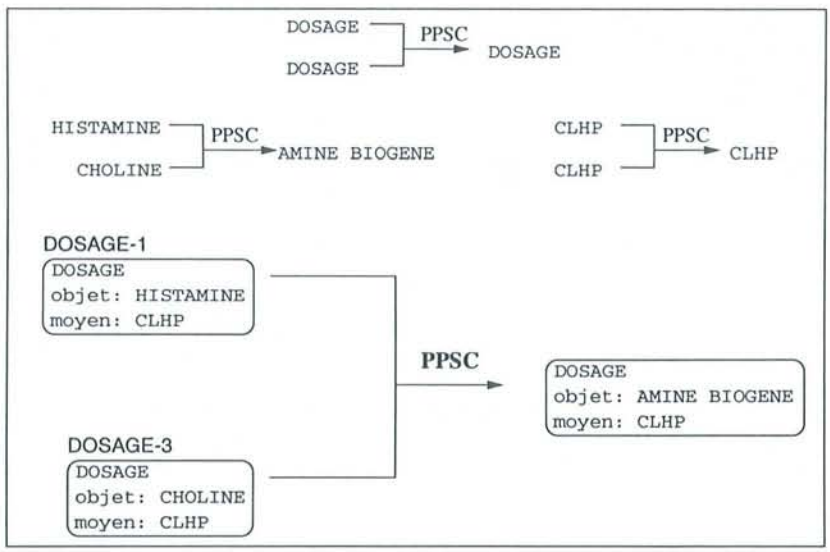

FIG. 4.1 – PPSC *de deux concepts et de chacune de leurs composantes*

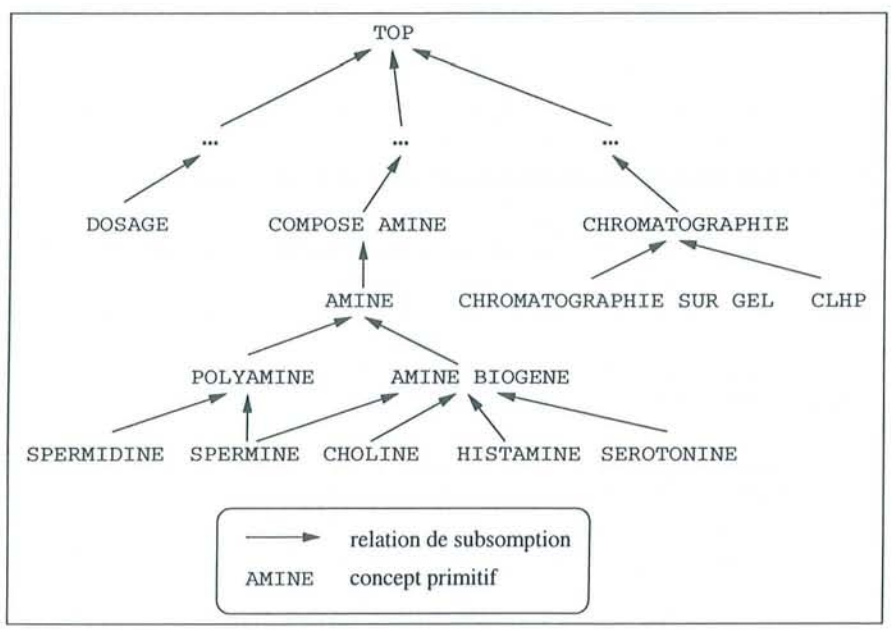

FIG. 4.2 – *Détail de la hiérarchie des concepts*

et DOSAGE-3 est la description conceptuelle exprimant la conjonction de la généralisation de chaque propriété des concepts. Puisque nous restreignons l'utilisation des opérateurs de CLASSIC, une propriété est exprimée soit par un nom de concept parent, par exemple,

DOSAGE dans DOSAGE-1,

soit par un rôle, par exemple,

`(all objet HISTAMINE)` dans `DOSAGE-1`.

Ainsi, dans notre exemple, le PPSC calculé possède une propriété mettant en jeu le rôle `objet` car celui-ci est présent dans les deux concepts à généraliser. La restriction du rôle `objet` du PPSC est le PPSC de la restriction du rôle `objet` de chacun des deux concepts `DOSAGE-1` et `DOSAGE-3`, de valeurs respectives `HISTAMINE` et `CHOLINE`, soit le concept `AMINE_BIOGENE`. Pour le rôle `moyen`, les deux restrictions sont identiques, égales à `CLHP`, la généralisation de cette propriété est donc triviale. Le PPSC de deux concepts primitifs dont la définition est limitée à la donnée de leur parent s'obtient par une remontée dans la hiérarchie des concepts, qui s'arrête dès qu'un subsumant commun est atteint. Dans le cas de parents multiples, le résultat peut être une conjonction de concepts, si ceux-ci sont incomparables selon la relation de subsomption.

Nous proposons un algorithme de calcul du PPSC de deux concepts qui se base sur les fonctions fournies par le système CLASSIC, contrairement aux algorithmes proposés par Cohen et al. ([Cohen 92], [Cohen 94b]) ou Ventos et al. ([Ventos 95]) qui reposent sur l'utilisation d'une structure de données appelée graphe de description, et nécessitent l'accès aux mécanismes internes d'une logique de descriptions. L'utilisation restreinte des opérateurs de construction de concepts, limitée à `and` et `all`, conduit à un algorithme simple et efficace.
L'algorithme de calcul du PPSC comprend trois fonctions : SUBSUMANT-PRIMITIF (algorithme 4) et LISTE-SUBSUMANTS (algorithme 5) permettent de calculer le PPSC de deux concepts primitifs sans rôles par une remontée de la hiérarchie des concepts. PPSC permet de calculer le PPSC de deux concepts quelconques de la hiérarchie (algorithme 6).

Les algorithmes utilisent des fonctions de base disponibles en CLASSIC ou en lisp, présentées figure 4.1. Dans l'ensemble des algorithmes présentés, nous utiliserons les notations suivantes pour désigner une table T dont les index sont $i_1, i_2, \ldots, i_n$ et les valeurs associées $v_1, v_2, \ldots, v_n$:

T représente la table entière
T(i_k) représente la valeur associée au k-ème index, soit v_k

Algorithme 4 calcul du plus petit subsumant commun de deux concepts primitifs sans rôles

SUBSUMANT-PRIMITIF(c_1, c_2)
- c_1 : concept primitif sans rôles
- c_2 : concept primitif sans rôles

si SUBSUME(c_2, c_1) **alors**
  retourne c_2
fin si
ListeConcepts ← BOTTOM(LISTE-SUBSUMANTS(c_1, c_2))
si TAILLE(ListeConcepts) > 1 **alors**
  retourne « (**and** + ListeConcepts +) »
sinon
  retourne PREMIER(ListeConcepts) *;;concept seul*
fin si

TAILLE(l : liste) : retourne le nombre d'éléments de la liste l
PREMIER(l : liste) : retourne le premier élément de la liste l
BOTTOM(lc : liste de concepts) : retourne la liste des concepts c_i de lc tels qu'aucun autre concept de la liste lc ne subsume c_i
DESCRIPTION(c : concept) : retourne la description en CLASSIC du concept c
RESTRICTION-ALL(r : rôle, c : concept) : retourne le concept qui représente le co-domaine du rôle r pour le concept c
LISTE-ROLES(c : concept) : retourne la liste des rôles utilisés pour dans la description du concept c
SUBSUME(c_1 : concept, c_2 : concept) : retourne VRAI si c_1 subsume c_2, FAUX sinon

FIG. 4.3 – *Fonctions utilisées dans les algorithmes*

Algorithme 5 calcul d'une liste de subsumants communs de deux concepts

LISTE-SUBSUMANTS(c_1, c_2)
- c_1 : concept primitif sans rôles
- c_2 : concept primitif sans rôles

si SUBSUME(c_1, c_2) **alors**
  retourne (c_1)
fin si
P $\leftarrow$ PARENTS(c_1)
ListeConcepts $\leftarrow \emptyset$
pour tout parent p $\in$ P **faire**
  ListeConcepts = ListeConcepts $\cup$ LISTE-SUBSUMANTS(p, c_2)
fin pour
retourne ListeConcepts

Les algorithmes 4 et 5 permettent de calculer le PPSC de deux concepts primitifs n'ayant pas de rôle, c'est-à-dire étant seulement défini par leur(s) parent(s). C'est le cas par exemple des concepts `HISTAMINE` et `CHOLINE` dont le PPSC est `AMINE_BIOGENE`. La fonction LISTE-SUBSUMANTS permet d'effectuer un parcours ascendant de la hiérarchie à partir d'un des deux concepts, jusqu'à trouver un concept qui subsume les deux concept initiaux. En cas de parenté multiple, chacun des parents est examiné. La liste finale obtenue par LISTE-SUBSUMANTS est ensuite filtrée par SUBSUMANT-PRIMITIF de façon à ne conserver que les subsumants les plus spécifiques (fonction BOTTOM). La description est complétée par le connecteur **and** dans le cas où plusieurs concepts sont possibles.

Par exemple, appliquons l'algorithme sur les concepts c_1 = `SPERMINE` et c_2 = `SEROTONINE`. LISTE-SUBSUMANTS retourne la liste {`AMINE_BIOGENE`, `AMINE`}. Celle-ci est réduite au singleton {`AMINE_BIOGENE`} par BOTTOM, le résultat final étant le concept `AMINE_BIOGENE`. Supposons qu'il existe une autre substance `X` qui soit comme `SPERMINE`, subsumée à la fois par `AMINE_BIOGENE` et `POLYAMINE`. Dans ce cas, pour c_2 = `X`, LISTE-SUBSUMANTS retourne la liste {`AMINE_BIOGENE`, `POLYAMINE`} et le PPSC est :

(**and** `AMINE_BIOGENE POLYAMINE`)

Algorithme 6 calcul du plus petit subsumant commun de deux concepts

PPSC(c_1, c_2)
- c_1 : concept
- c_2 : concept

si SUBSUME(c_1, c_2) **alors**
  retourne DESCRIPTION(c_1)
fin si
si SUBSUME(c_2, c_1) **alors**
  retourne DESCRIPTION(c_2)
fin si
Genre ← SUBSUMANT-PRIMITIF(c_1, c_2)
L_{inter} ← LISTE-ROLES(c_1) ∩ LISTE-ROLES(c_2)
Description ← ∅
pour tout role r ∈ L_{inter} **faire**
  $rest_1$ ← RESTRICTION-ALL(r, c_1)
  $rest_2$ ← RESTRICTION-ALL(r, c_2)
  DescPartielle ← « (**all** + r + PPSC($rest_1, rest_2$) +) »
  Description ← Description + DescPartielle
fin pour
si Description = ∅ **alors**
  retourne Genre
sinon
  retourne « (**and** + Genre + Description +) »
fin si

L'algorithme 6 permet de calculer le PPSC pour deux concepts quelconques (primitifs ou définis). Après avoir vérifié qu'aucun des deux concepts ne subsume l'autre, le PPSC est

calculé en décomposant les propriétés des deux concepts, rôle par rôle, en appliquant récursivement l'algorithme. Si un rôle n'apparaît que dans un des deux concepts, il sera absent du PPSC. L'application de SUBSUMANT-PRIMITIF permet de calculer la « composante parentale », c'est-à-dire le ou les parents du plus petit subsumant commun.

Le calcul du PPSC n'est pas limité aux structures prédicatives simples. Il permet également de prendre en compte les structures complexes dans lesquelles plusieurs structures prédicatives sont imbriquées. Appliquons l'algorithme sur un exemple faisant intervenir une description complexe. Soient les deux concepts suivants :

```
SIMULATION-1 ≐ (and SIMULATION
                    (all objet (and CROISSANCE
                                    (all objet MAIS))))

SIMULATION-2 ≐ (and SIMULATION
                    (all objet (and DEVELOPPEMENT
                                    (all objet MAIS))))
```

SIMULATION-1 et SIMULATION-2 représentent respectivement les groupes nominaux suivants : *simulation de la croissance du mais* et *simulation du développement du mais.* Le PPSC des concepts CROISSANCE et DEVELOPPEMENT est le concept EVOLUTION. L'application de l'algorithme PPSC sur SIMULATION-1 et SIMULATION-2 conduit au déroulement suivant :

Genre = SIMULATION
L_{inter} = {objet}

Le traitement du rôle objet consiste à appliquer récursivement PPSC sur les deux restrictions :

(and CROISSANCE **(all** objet MAIS**)**
et
(and DEVELOPPEMENT **(all** objet MAIS**))**

Finalement, le PPSC est décrit par :

```
G1-SIMULATION ≐ (and SIMULATION
                     (all objet (and EVOLUTION
                                     (all objet MAIS))))
```

L'algorithme de calcul du PPSC permet ainsi de traiter toutes les structures prédicatives que l'on peut exprimer à l'aide des opérateurs **and** et **all**, même lorsque les rôles sont complexes, c'est-à-dire lorsqu'ils contiennent une structure prédicative interne. Nous montrons dans la section suivante comment nous utilisons cet opérateur pour introduire des généralisations dans la hiérarchie des concepts.

4.2 Introduction d'une généralisation en CLASSIC

Nous savons maintenant comment calculer le PPSC d'un ensemble C de concepts. Le PPSC obtenu est une description de concept, qui généralise les structures prédicatives représentées par les concepts de l'ensemble C, et que nous appelons *généralisation* pour faire

une distinction avec les PPSC calculables sur tous les ensembles possibles. Pour introduire de manière effective cette généralisation dans la hiérarchie des concepts, il faut la nommer et décider quel statut lui donner :

- par convention, une généralisation sera nommée par un « G » suivi d'un nombre unique et du nom de la tête prédicative la composant,
- une généralisation est introduite comme *concept défini*, ceci afin d'exploiter le mécanisme de classification (déductif) de la logique de descriptions, qui permet une mise à jour automatique de la hiérarchie des concepts.

Reprenons l'exemple des sept structures prédicatives introduites en 3.2, avec la tête prédicative *dosage*. Nous avons calculé le PPSC des deux concepts DOSAGE-1 et DOSAGE-3 (figure 4.1, page 54) dont la description est la suivante :

```
(and DOSAGE
     (all objet AMINE_BIOGENE)
     (all moyen CLHP))
```

Si nous voulons introduire cette généralisation en CLASSIC, nous le faisons à l'aide d'un concept défini :

```
G1-DOSAGE ≐ (and DOSAGE
                 (all objet AMINE_BIOGENE)
                 (all moyen CLHP))
```

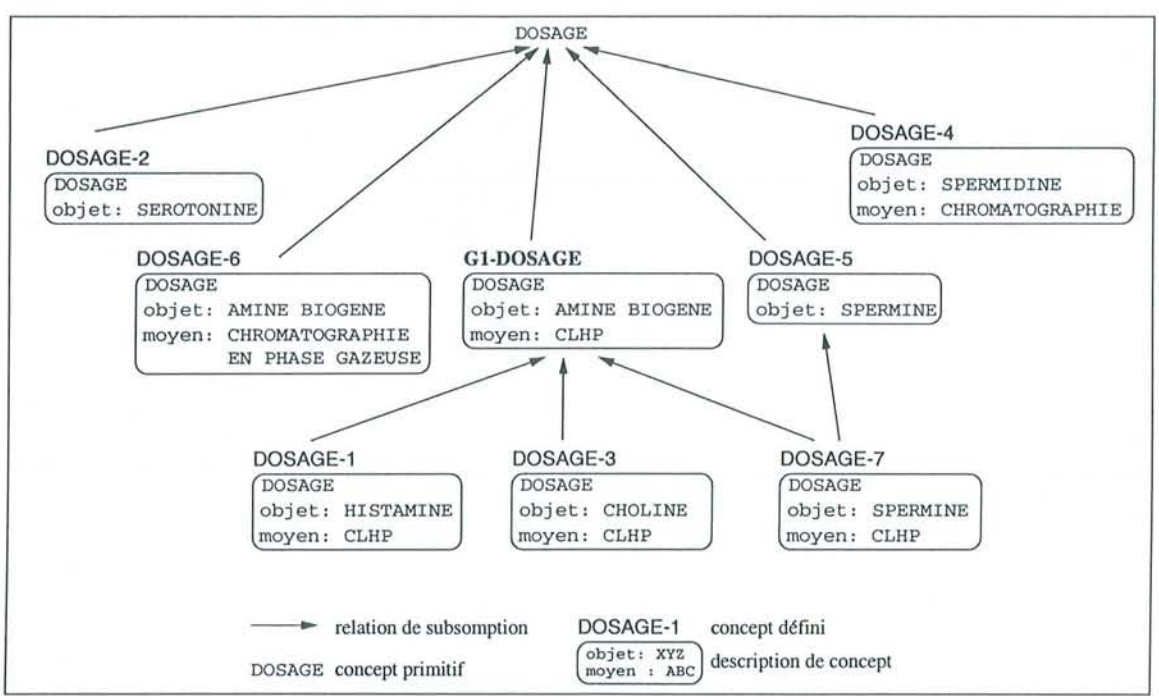

FIG. 4.4 – *Hiérarchie conceptuelle avec l'introduction de la généralisation G1-DOSAGE*

La hiérarchie résultante, après l'introduction de G1-DOSAGE, est illustrée par la figure 4.4. G1-DOSAGE est par définition plus général que les concepts qui ont contribué à la générer, DOSAGE-1 et DOSAGE-3. D'autre part, DOSAGE-7 est plus spécifique que G1-DOSAGE est se

trouve donc placé également sous ce concept. Le lien existant entre DOSAGE-7 et DOSAGE-5 est conservé car DOSAGE-5 est incomparable avec G1-DOSAGE, c'est-à-dire ni plus général, ni plus spécifique. On observe ainsi que l'introduction d'une généralisation en CLASSIC conduit à réorganiser la hiérarchie par la calcul des nouveaux liens de subsomption : la généralisation G1-DOSAGE a été calculée à partir de deux concepts, mais en recouvre trois.

Nous pouvons insister ici sur les avantages de l'utilisation d'une logique de descriptions pour la généralisation. Le mécanisme déductif de classification des concepts offre un cadre logique pour comparer et classer les structures prédicatives. Les comparaisons sont effectuées selon des critères bien définis, ceux de la subsomption. Le caractère inductif de la généralisation est assuré par l'utilisation de l'opération inductive du PPSC, qui permet de calculer des concepts plus généraux. Nous disposons ainsi d'un cadre pour la classification conceptuelle qui est bien défini logiquement, opérationnel et efficace.

Nous disposons à présent des éléments pour proposer un processus global de généralisation. Cependant, l'application de l'opération de PPSC doit être contrainte pour permettre d'aboutir à des généralisations en nombre limité et pertinentes. Nous présentons et motivons dans la section suivantes les principes que nous avons choisis d'appliquer pour atteindre cet objectif.

4.3 Des principes et des heuristiques pour la généralisation

Étant donné un ensemble C de n concepts, représentants des structures prédicatives, chaque élément de $\mathcal{P}(C)$, l'ensemble des parties de C, peut être candidat à la généralisation par application de l'opération de PPSC. Un processus de généralisation exhaustif consisterait donc à calculer le PPSC de chaque sous-ensemble de $\mathcal{P}(C)$, et à introduire la généralisation correspondante en CLASSIC. Cette approche est cependant peu souhaitable pour deux raisons principales :

- le nombre de généralisations à calculer est de croissance exponentielle suivant n et conduit à une complexité calculatoire très élevée,
- le nombre de généralisations obtenues est très élevé (malgré une probable redondance), et une proportion importante de ces généralisations est de « faible qualité », c'est-à-dire difficilement interprétable par l'utilisateur final.

Il s'avère donc indispensable de trouver des moyens pour contourner la complexité calculatoire inhérente au processus tout en favorisant les généralisations susceptibles d'être de « bonne qualité ». Cette qualité est évidemment une notion subjective qui dépend fortement de la nature des informations traitées et de l'objectif final visé. Nous avons montré dans le chapitre 2 que l'« approche classique » de la classification conceptuelle, par exemple, n'est pas satisfaisante car elle s'intéresse surtout à la prédiction de valeurs inconnues. Dans notre cas, il s'agit de retenir des généralisations qui soient synthétiques sans être trop générales, et qui soient surtout intelligibles pour l'utilisateur, en rapportant des informations contenues dans les textes. Les informations doivent être regroupées en tenant compte de leur présentation sous forme de structures prédicatives. Par exemple, la généralisation de deux structures prédicatives n'ayant aucun rôle en commun donne une structure prédicative très peu informative limitée à un concept primitif représentant une tête prédicative : il est donc inutile d'effectuer l'opération de calcul du PPSC.

4.3.1 Utilisation de la hiérarchie comme indice de proximité des structures prédicatives

Toutes les structures prédicatives n'entretiennent pas le même rapport. Certaines sont plus proches que d'autres en qui concerne leur contenu informatif, et susceptibles d'être rapprochées, c'est-à-dire d'être généralisées. Il s'agit donc de juger la similarité des structures prédicatives. Étant donnée la forme d'une structure prédicative, deux types d'informations sont comparables :

- la tête prédicative utilisée par chaque structure prédicative,
- les couples relation-argument utilisés par chaque structure prédicative.

Dans les deux cas, nous utilisons l'ordre partiel imposé par la hiérarchie des concepts comme base pour la comparaison des concepts (têtes prédicatives et arguments). Les concepts sont donc regroupés selon leur proximité dans la hiérarchie. Ceci permet par la suite de regrouper les structures prédicatives qui doivent être généralisées ensembles.

Par exemple, considérons une portion de la hiérarchie des concepts, illustrée par la figure 4.5. La comparaison des concepts `SPERMIDINE`, `CHOLINE` et `HISTAMINE`, conduit à regrouper les deux concepts `CHOLINE` et `HISTAMINE` : en effet, ils sont plus proches en terme d'information, puisque leur généralisation conduit au concept `AMINE_BIOGENE` qui ne recouvre pas le concept `SPERMIDINE`. En revanche, la comparaison des concepts `AMINE_BIOGENE`, `POLYAMINE` et `HYDROXYLAMINE` conduit à un seul regroupement contenant les trois concepts, car ceux-ci ont une position identique dans la hiérarchie. Leur généralisation est représentée par le concept `AMINE`. Ces exemples illustrent l'importance que nous donnons à la hiérarchie pour regrouper les concepts, et, par extension, les structures prédicatives pour la généralisation.

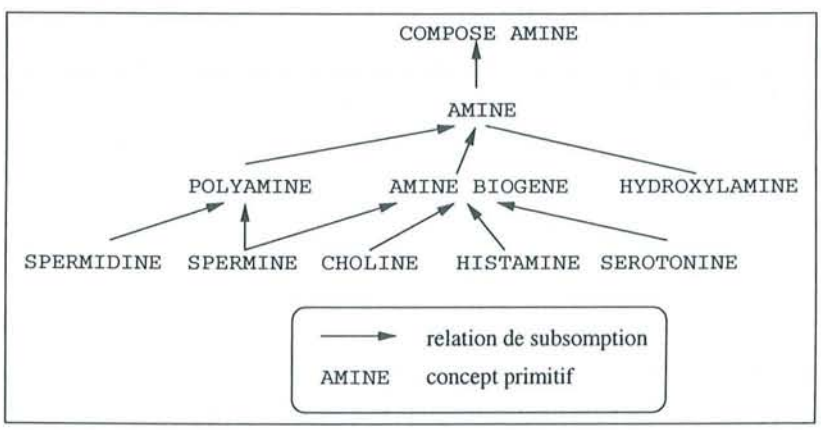

FIG. 4.5 – *Détail de la hiérarchie des concepts*

4.3.2 Une heuristique qui décompose le problème de généralisation en deux étapes

Les structures prédicatives peuvent être regroupées en utilisant la proximité des têtes prédicatives et la proximité des arguments. Nous pensons que la similarité des têtes prédicatives est plus importante que celle des arguments pour générer une structure prédicative

pertinente. En effet, deux structures prédicatives partageant le même prédicat sont toujours de sens « voisin », même lorsque les arguments diffèrent : le prédicat tel que nous l'avons défini nous assure une proximité conceptuelle minimum entre les deux structures. Au contraire, deux structures prédicatives partageant les mêmes arguments mais ayant deux prédicats différents peuvent être très éloignées conceptuellement l'une de l'autre. Ceci est d'autant plus fréquent que le nombre d'arguments est petit.

Par exemple, considérons les trois structures prédicatives suivantes :

1 : *dosage(objet : choline)*
2 : *dosage(objet : spermidine)*
3 : *production(objet : choline)*

Les structures prédicatives *1* et *2* doivent être regroupées en priorité. Le regroupement des structures *1* et *3*, par ailleurs intéressant, est moins pertinent car il conduit à une généralisation dont la tête prédicative est très générale et l'argument très spécifique.
Cette constatation nous amène à proposer une première heuristique qui consiste à décomposer le processus de généralisation en deux étapes distinctes :

- une première étape où chaque groupe de structures prédicatives partageant le même prédicat est traité isolément,
- une deuxième étape où les informations généralisées pour chaque prédicat sont mises en commun.

Cette décomposition permet de ne garder pour la seconde étape que des informations assez synthétiques, et de donner moins d'importance aux informations marginales relatives à un prédicat donné. Ainsi, de nombreux calculs et généralisations peu pertinents sont évités.

Par exemple, considérons un ensemble E de vingt structures prédicatives : dix structures prédicatives possèdent la tête prédicative *dosage*, six structures possèdent la tête prédicative *identification*, trois structures possèdent la tête prédicative *analyse* et une structure possède le prédicat *production*. La première étape du processus de généralisation consiste à considérer quatre ensembles D, I, A et P comprenant respectivement les structures prédicatives dont la tête prédicative est *dosage*, *identification*, *analyse* et *production*. Pour chacun de ces ensembles pris séparément, des généralisations sont calculées en appliquant l'opération de calcul du PPSC sur des sous-ensembles choisis selon la proximité des structures prédicatives (ce processus est détaillé section 4.4).

La première étape permet d'aboutir à quatre ensembles de structures prédicatives D', I', A' et P', sur-ensembles respectivement de D, I, A et P. Par exemple, I' est l'ensemble I auquel sont ajoutées les généralisations calculées sur I. Toutes les structures prédicatives de l'ensemble I' possèdent la tête prédicative *identification* car les généralisations ajoutées à I' sont issues de structures prédicatives dont la tête prédicative est *identification*.

La deuxième étape consiste à ne retenir, dans les ensembles D', I', A' et P' contenant les généralisations calculées, qu'un petit nombre de structures prédicatives représentatives des informations contenues dans l'ensemble considéré. Celles-ci sont alors regroupées principalement selon la position hiérarchique de leur tête prédicatives : la position relative des concepts utilisés comme tête prédicative est donc le principal critère de regroupement pour cette deuxième étape (présentée en détail section 4.5).

Cette approche pose comme hypothèse la possibilité d'établir une hiérarchie de prédicats suffisamment pertinente et intelligible. C'est-à-dire qu'il doit être possible, pour le

domaine de connaissances considéré, d'établir un ordre partiel non trivial sur les différentes actions ou événements décrivant le domaine. Une telle hiérarchie a par exemple été établie pour la langue générale, par les travaux sur Wordnet, une base de connaissances lexicales en anglais [Miller 93]. Elle montre qu'une telle approche est possible. Nous considérons qu'elle est plus appropriée pour une langue de spécialité, plus contrainte, que pour la langue générale, soumise à une plus grande polysémie[35]. Nous pensons toutefois qu'elle doit se limiter à un domaine restreint, voire un sous-domaine, pour être exploitable. Nous donnons figure 4.6 un extrait d'ordre partiel sur des mots prédicatifs (le plus souvent des verbes) établi par la base lexicale Wordnet : un noeud correspond à une notion, et s'exprime par un ensemble de mots qui permettent de cerner la notion considérée. La relation d'ordre entre les noeuds est nommée troponymie (*troponomy* en anglais) par les auteurs, et signifie intuitivement : « est une façon particulière de ». Nous précisons entre parenthèses le sens approximatif de ces notions en français, au moyen d'un mot ou d'un ensemble de mots. L'exploitation d'une telle hiérarchie s'avère particulièrement ardue à cause de la grande polysémie des mots utilisés. Par exemple, le mot *analyse* possède trois sens dans Wordnet, dont un seul est illustré par la figure 4.6.
Se limiter à un domaine restreint est donc un moyen de réduire le problème de la polysémie. D'autre part, pour pouvoir être exploitée efficacement par l'utilisateur final, la hiérarchie doit être construite de manière consensuelle et doit impliquer l'utilisateur. Cette question sera approfondie dans le chapitre concernant l'évaluation du processus de généralisation (chapitre 6).
Par la suite, nous ferons référence à cette heuristique de décomposition de la généralisation

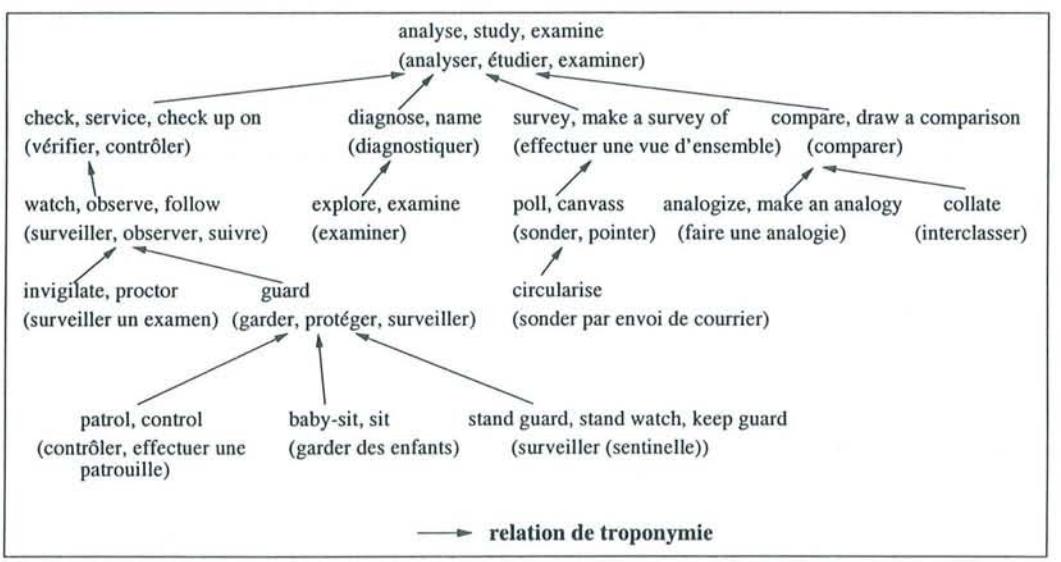

FIG. 4.6 – *Un extrait de la hiérarchie des verbes de Wordnet, avec leur traduction en français*

en deux étapes par l'appellation *heuristique* $\mathcal{H}_1$.

35. Un mot est dit *polysémique* lorsqu'il renvoie à plusieurs notions.

4.3.3 Une heuristique pour traiter le problème de la diversité des structures prédicatives

Nous avons déjà observé que l'arité des structures prédicatives peut être sujette à des variations, et que les structures prédicatives susceptibles d'être extraites des textes peuvent être incomplètes (cf. section 3.2). Nous sommes donc confrontés pour la généralisation à des structures prédicatives qui peuvent avoir un nombre de relations différent et des relations différentes, même si elles partagent la même tête prédicative. Si deux structures prédicatives n'ont aucune relation commune, leur généralisation est inutile puisqu'elle ne peut donner qu'un concept réduit à un prédicat, peu intéressant : nous pouvons donc réduire le nombre de généralisations obtenues en interdisant de regrouper de telles structures prédicatives. Pour les autres cas de figure, où des structures prédicatives possèdent en partie seulement des relations communes, il existe plusieurs approches possibles, dont les deux extrêmes consistent soit à autoriser tous les types de regroupements soit au contraire à les limiter au maximum.

La question que nous nous posons est donc la suivante : sur quels critères deux (et par extension n) structures prédicatives peuvent elles être généralisées ensemble? Soient deux structures prédicatives sp_1 et sp_2, possédant respectivement les relations $r_{1i}, 1 \leq i \leq n$ et $r_{2i}, 1 \leq i \leq m$, nous énumérons différents critères possibles : la généralisation de sp_1 et sp_2 est autorisée si

1 il existe au moins une valeur i et un valeur j tel que $r_{1i} = r_{2j}$, c'est-à-dire s'il existe au moins une relation en commun, par exemple :
dosage(objet : amine, moyen : chromatographie liquide, localisation : vin)
dosage(objet : polyamine),

2 il existe au moins $max(m,n)-1$ valeurs i et $max(m,n)-1$ valeurs j tel que $r_{1i} = r_{2j}$, c'est-à-dire s'il existe au plus une relation qui n'est pas en commun dans chaque structure prédicative, par exemple :
dosage(objet : amine, moyen : chromatographie liquide)
dosage(objet : polyamine, localisation : farine de poisson),

3 il existe au moins k valeurs i et k valeurs j tel que $r_{1i} = r_{2j}$, c'est-à-dire s'il existe au moins k relations en commun (le cas 1 est un cas particulier de celui-ci), par exemple (pour $k = 2$) :
dosage(objet : amine, moyen : chromatographie liquide)
dosage(objet : polyamine, moyen : clhp, localisation : farine de poisson),

4 $m = n$ et il existe j tel que $r_{1i} = r_{2j}$ pour tout i vérifiant $1 \leq i \leq m$, c'est-à-dire si toutes les relations sont en commun, par exemple :
dosage(objet : amine, moyen : chromatographie liquide)
dosage(objet : polyamine, moyen : clhp).

La solution *1* consiste à autoriser la généralisation d'un ensemble de structures prédicatives si celles-ci possèdent au moins une relation en commun : cela signifie que deux structures prédicatives ayant respectivement les relations *objet*, *localisation* et *objet*, *moyen* sont généralisables. Elles conduisent donc à une structure prédicative ne possédant que la relation *objet* : les informations initiales sont fortement appauvries, et la généralisation résultante a une forte probabilité de subsumer un nombre beaucoup plus important de structures prédicatives que le nombre initial qui a contribué à son calcul. De plus, cette solution est celle qui conduit au nombre maximal de généralisations possibles.

La solution *4* est à l'opposé de la précédente, et consiste à imposer la correspondance exacte entre les relations possédées par un ensemble de structures prédicatives pour calculer leur généralisation : cela permet d'aboutir à des généralisations qui possèdent autant de relations que les structures prédicatives qui ont contribuées à leur calcul. Les informations initiales sont donc appauvries (ou, d'un autre point de vue, synthétisées) au niveau des arguments seulement. Cette solution est également la plus économique car elle supprime un nombre important de généralisations potentielles, et réduit la complexité calculatoire.

Les solutions *2* et *3* sont des compromis possibles entre les deux solutions extrêmes *1* et *4*. Cependant, elles sont plus arbitraires que les précédentes, car il est difficile d'établir *a priori* une limite sur le nombre de relations en commun conduisant à une généralisation « acceptable ». Cette limite pourrait être établie à partir d'expérimentations sur divers corpus, mais il semble très peu probable qu'elle puisse être fixée de manière générale.

Nous retrouvons ici le compromis entre la qualité des généralisations, la complexité du processus et la couverture des généralisations : l'approche de classification conceptuelle utilisant les treillis (section 2.4) ou les graphes conceptuels (section 2.5) consiste à générer tous les concepts possibles au détriment de la complexité et de la qualité des généralisations. Toutefois, l'objectif visé par Simon et Napoli [Simon 98] ou Bournaud [Bournaud 96] permet de tolérer une hiérarchie complexe, celle-ci étant explorée et simplifiée patiemment par l'utilisateur. En effet, celui-ci tente de repérer des classifications intéressantes en utilisant les mécanismes de simplifications de hiérarchies proposés par les auteurs. Dans notre cas, l'utilisateur ne doit pas passer du temps à tester diverses configurations et diverses hiérarchies car il désire accéder au contenu de manière immédiate.
Dans la mesure où l'objectif est de faciliter l'accès du contenu informationnel du corpus, il n'est pas souhaitable de surcharger l'utilisateur avec des tâches annexes. Ce dernier doit pouvoir profiter de l'organisation hiérarchique sans avoir à la modifier ou la simplifier. En cela, nous sommes proche de l'objectif de Carpineto et Romano [Carpineto 96], qui génèrent une hiérarchie de thèmes pour la recherche d'information et ne peuvent demander à l'utilisateur de modifier la hiérarchie. Cependant, leur travail utilise des expressions linguistiques simples, limitées à des termes, c'est-à-dire des mots ou groupes de mots. Dans notre cas, l'utilisation de structures prédicatives contribue à une complexité accrue de l'interprétation de la hiérarchie. Pour cette raison, nous choisissons de limiter le nombre de généralisations qu'il est possible de générer, en adoptant la solution *4*, c'est-à-dire en imposant une correspondance exacte entre les relations des structures prédicatives pour qu'elles puissent être généralisées ensembles. Ce choix constitue notre *deuxième heuristique*, $\mathcal{H}_2$.

Illustrons ces propos par un exemple : considérons les sept structures prédicatives `DOSAGE-1`, `DOSAGE-2`, ..., `DOSAGE-7` introduites précédemment (la figure 4.7 reproduit la hiérarchie correspondante). Le choix que nous avons effectué nous conduit à limiter les généralisations possibles. Ainsi, les concepts `DOSAGE-1`, `DOSAGE-3`, `DOSAGE-4`, `DOSAGE-6` et `DOSAGE-7` peuvent être généralisés ensemble car ils possèdent le même ensemble de rôles, {`objet`, `moyen`}. Indépendamment, les concepts `DOSAGE-2` et `DOSAGE-5` peuvent faire l'objet d'une généralisation. Si un concept `DOSAGE-8` possédant l'ensemble de rôle {`moyen`} était également considéré, il ne pourrait être généralisé avec les autres concepts. Dans la pratique, une telle variété des ensembles de rôles est peu souhaitable, car elle contribue à la difficulté de l'interprétation de la hiérarchie. Nous faisons l'hypothèse que les ensembles de rôles sont relativement homogènes pour un prédicat donné.

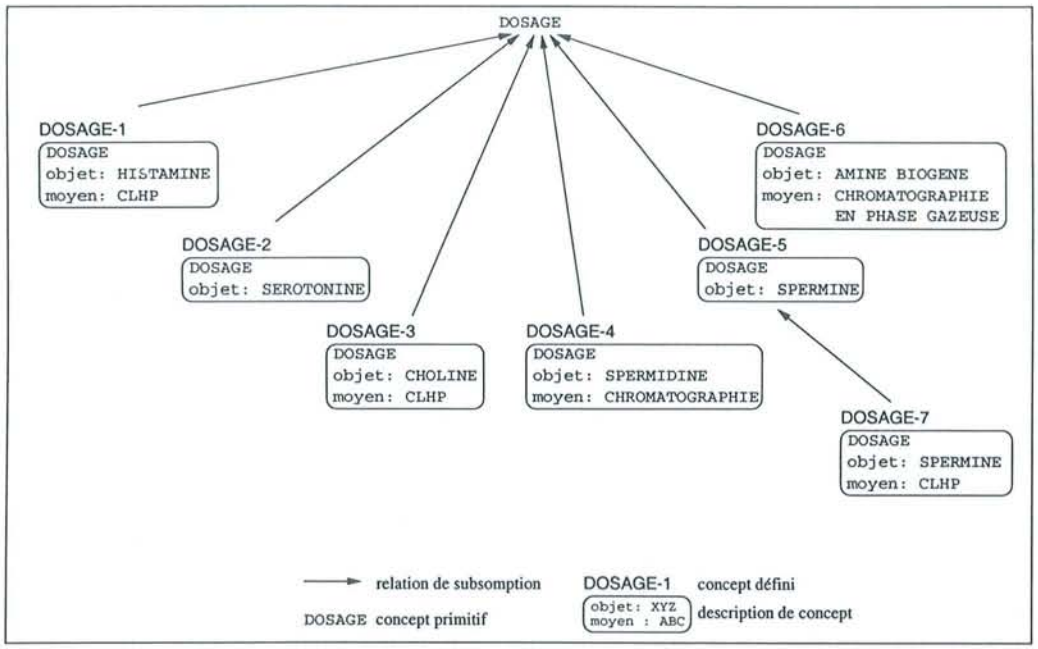

FIG. 4.7 – *Hiérarchie de sept structures prédicatives* `DOSAGE-1`, ..., `DOSAGE-7` *en* CLASSIC

Conclusion

Nous avons énoncé les principes et heuristiques suivantes pour la généralisation des structures prédicatives :

- les structures prédicatives sont regroupées selon leur « proximité sémantique », donnée par l'ordre défini par la hiérarchie conceptuelle,
- la généralisation est décomposée en deux étapes, la première permettant une généralisation selon la tête prédicative, et la deuxième permettant la mise en commun de l'ensemble des structures prédicatives (heuristique $\mathcal{H}_1$),
- les structures prédicatives regroupées pour conduire à une généralisation doivent posséder le même ensemble de rôles (heuristique $\mathcal{H}_2$).

Nous présentons en détail les deux étapes du processus de généralisation dans les deux section suivantes.

4.4 Première étape de la généralisation : prédicat par prédicat

Dans cette section, nous considérons un ensemble SP de n structures prédicatives possédant toutes la même tête prédicative. L'objectif est de calculer des généralisations de ces structures à l'aide de l'opération de calcul de PPSC. Nous présentons tout d'abord de manière informelle chaque étape du processus, suivi de l'algorithme correspondant, puis de son application sur un exemple. L'exemple utilisé tout au long de cette section est l'ensemble de structures prédicatives `DOSAGE-1`, `DOSAGE-2`, ..., `DOSAGE-7`, déjà introduit au chapitre précédent (cf. page 48 et suivantes).

Dans un premier temps, les structures prédicatives sont séparés en plusieurs sous-ensembles, selon les rôles qu'elles possèdent, pour respecter l'heuristique $\mathcal{H}_2$ présenté précédemment (section 4.3.3). Ainsi, il peut y avoir un sous-ensemble de structures prédicatives possédant l'ensemble de rôles {`objet`, `moyen`}, un autre possédant l'ensemble de rôles {`objet`}, et ainsi de suite. Pour chacun des sous-ensembles, nous calculons la généralisation de toutes les structures prédicatives le composant, obtenant ainsi la structure prédicative la plus générale possible pour le sous-ensemble. Elle est calculée en premier car elle est utilisée par la suite pour fixer une borne supérieure (au sens de la relation de subsomption) aux arguments des structures prédicatives. Ce premier traitement est effectué par la fonction DECOMPOSER-PREDICAT donnée par l'algorithme 7. Dans notre exemple, la

Algorithme 7 Traitement des structures prédicatives possédant le prédicat p

DECOMPOSER-PREDICAT(p, LSP)
- p : concept représentant un prédicat
- LSP : liste de concepts représentant des structures prédicatives

TRoles ← nil ;; *TRole est une table*
pour chaque structure prédicative sp_i de LSP **faire**
  ER_i ← LISTE-ROLES(sp_i)
  TRoles(ER_i) ← TRoles(ER_i) ∪ sp_i
fin pour
pour chaque ensemble de rôles ER_i de TRoles **faire**
  si TAILLE(TRoles(ER_i)) >1 **alors**
    g ← PPSC(Table(ER_i)) ;; *Table(ER_i) =(sp_1, sp_2, ..., sp_n)*
    introduire g en CLASSIC et le marquer comme généralisation
    GENERALISER-ENSEMBLE-DE-ROLES(g, TRoles(ER_i))
  fin si
fin pour

table *TRoles* possède deux entrées : l'ensemble de rôles {`objet`, `moyen`} et l'ensemble de rôles {`objet`}. Ainsi deux sous-ensembles sont crées, SP_1 et SP_2 :

pour l'ensemble {`objet`, `moyen`}
→ SP_1 : {`DOSAGE-1`, `DOSAGE-3`, `DOSAGE-4`, `DOSAGE-6`, `DOSAGE-7`}
pour l'ensemble {`objet`}
→ SP_2 : {`DOSAGE-2`, `DOSAGE-5`}

Une généralisation est calculée sur chaque sous-ensemble :

pour SP_1 : PPSC({`DOSAGE-1`, `DOSAGE-3`, `DOSAGE-4`, `DOSAGE-6`, `DOSAGE-7`}) →

```
G1-DOSAGE ≐ (and DOSAGE
                 (all objet AMINE)
                 (all moyen CHROMATOGRAPHIE))
```

pour SP_2 : PPSC({`DOSAGE-2`, `DOSAGE-5`}) →

```
G2-DOSAGE ≐ (and DOSAGE
                 (all objet AMINE_BIOGENE)
```

Ces généralisations permettent de fixer une « borne supérieure » aux arguments : par exemple, pour toutes les structures prédicatives de SP_1, les arguments liés au rôle *objet* seront plus spécifiques que le concept AMINE. Cette information est utilisée par la suite pour restreindre la recherche des subsumants des arguments lors de leur généralisation.

L'étape suivante consiste à trouver pour chaque sous-ensemble SP_i des généralisations plus spécifiques, intermédiaires entre les structures prédicatives initiales et la généralisation calculée sur le sous-ensemble complet.

Pour chaque sous-ensemble SP_i, les structures prédicatives diffèrent uniquement par les arguments associés à chaque rôle. Comme nous l'avons expliqué à la section précédente, l'ordre partiel sur les concepts permet de regrouper les structures prédicatives selon la généralité de leurs arguments. Toutefois, dans le cas où plusieurs rôles sont en jeu, deux attitudes sont possibles :

- considérer chaque rôle indépendamment,
- considérer les rôles ensembles, selon une heuristique permettant de simuler un ordre de généralité.

Considérons par exemple, deux structures prédicatives sp_1 et sp_2 possédant les rôles suivants :

sp_1 : `(all objet SEROTONINE) (all moyen CHROMATOGRAPHIE)`
sp_2 : `(all objet AMINE) (all moyen CLHP)`

Si l'on considère les rôles séparément, il est facile d'ordonner les structures prédicatives selon la généralité de leur argument pour un rôle donné. Ainsi, comme l'argument SEROTONINE est plus spécifique que l'argument AMINE, la structure prédicative sp_1 possédant l'argument SEROTONINE est considérée comme plus spécifique que la structure prédicative sp_2 pour le rôle objet. Par contre, les considérer ensemble nécessite de définir une combinaison de la généralité des arguments, qui ne peut être qu'approximative : dans l'exemple, les arguments étant inversement ordonnés selon le rôle choisi, deux méthodes différentes peuvent aboutir à deux solutions différentes. En effet, sp_1 est plus spécifique que sp_2 pour le rôle objet et plus générale que sp_2 pour le rôle moyen.
Nous préférons donc choisir la première solution, qui s'avère plus exhaustive et repose sur une opération logique, la subsomption. Bien sûr, ce choix a un coût en terme de calculs, puisque traiter chaque rôle séparément multiplie le nombre d'opérations par le nombre de rôles possédé par chaque structure prédicative. Toutefois, en pratique, ce nombre reste petit, puisqu'il dépasse rarement quatre, les structures prédicatives ayant généralement deux à trois arguments.

Pour mettre en oeuvre la généralisation des structures prédicatives intermédiaires, nous traitons chaque rôle séparément, et calculons la liste des concepts qui sont des restrictions du rôle. Cette liste est utilisée ensuite pour regrouper les structures prédicatives et les généraliser selon la spécificité des restrictions. L'algorithme 8 (fonction GENERALISER-ENSEMBLE-DE-ROLES) détaille ce traitement. Sur notre exemple, considérons le sous-ensemble de structures prédicatives SP_1. Il y a deux rôles à traiter, objet et moyen. Les listes de restrictions respectives sont :

$LREST_{objet}$ = {AMINE_BIOGENE, CHOLINE, HISTAMINE, SPERMIDINE, SPERMINE}
$LREST_{moyen}$ = {CLHP, CHROMATOGRAPHIE, CHROMATOGRAPHIE_EN_PHASE_GAZEUSE}

Algorithme 8 Traitement des structures prédicatives possédant le même ensemble de rôles

GENERALISER-ENSEMBLE-DE-ROLES(c, LSP)
- c : concept représentant une structure prédicative
- LSP : liste de concepts représentant des structures prédicatives

$LR_c \leftarrow$ LISTE-ROLES(c)
pour chaque role r de LR_c **faire**
  LREST $\leftarrow$ 0, Assoc $\leftarrow$ 0 *;; Assoc est une table associant une structure prédicative à sa restriction sur le rôle considéré*
  pour chaque structure prédicative sp_i de LSP **faire**
    $rest_i \leftarrow$ RESTRICTION-ALL(sp_i, r)
    LREST $\leftarrow$ LREST $\cup$ $rest_i$
    Assoc($rest_i$) $\leftarrow$ Assoc($rest_i$) $\cup$ sp_i
  fin pour
  PARCOURIR(RESTRICTION-ALL(c, r), LREST)
fin pour

La table `Assoc` est utilisée pour retrouver par la suite la structure prédicative correspondant à une restriction. Par exemple, pour le rôle `moyen`, la valeur de Assoc(`CLHP`) est {`DOSAGE-1`, `DOSAGE-3`, `DOSAGE-7`}. Les listes de restrictions *LREST* sont considérées comme des ensembles, par conséquent les restrictions redondantes sont supprimées.

A partir de la hiérarchie des concepts, nous construisons une sous-hiérarchie limitée aux concepts contenus dans la liste des restrictions et à leurs ascendants dans la hiérarchie des concepts (algorithme 9). Cette sous-hiérarchie est représentée par un graphe orienté distinct du graphe représentant la hiérarchies des concepts. Le parcours récursif descendant du graphe construit permet ensuite de calculer progressivement les généralisations lors de la remontée (algorithmes 10 et 11).

Algorithme 9 Construction du graphe orienté à partir d'une liste de restrictions et de leur ascendants

PARCOURIR(ctop, LC)
- ctop : concept représentant la restriction maximale sur un rôle
- LC : liste de concepts représentant les arguments des relations

pour chaque concept c_i de LC **faire**
  $L_{ASC}(c_i) \leftarrow$ liste des ascendants de c_i plus spécifiques que ctop
fin pour
$L_{DESC} \leftarrow$ liste inverse de L_{ASC}
DEBUT-PARCOURS(ctop)

Dans l'algorithme 9, les tables L_{ASC} et L_{DESC} sont les structures de données permettant de représenter le graphe. Ce sont des listes d'adjacence particulières : L_{ASC} contient la liste de tous les ascendants de chaque restriction qui sont plus spécifiques que *ctop* ; L_{DESC} est la liste inverse, et associe à chaque concept ayant un descendant son ou ses descendants immédiats. Ainsi, les concepts indexés par L_{ASC} coïncident avec les concepts

qui sont des restrictions. Et les concepts indexés par L_{DESC} coïncident avec les concepts intermédiaires entre la borne supérieure (*ctop*, égale à la restriction de la généralisation maximale calculée sur le sous-ensemble traité) et les concepts restrictions (les concepts intermédiaires pouvant également être des concepts restrictions).

Pour illustrer ces propos, nous donnons, sur la figure 4.8, le graphe calculé pour le traitement du rôle *objet* pour le sous-ensemble SP_1. Dans ce cas, le concept *ctop* des algorithmes 9 et 10 est le concept AMINE, qui représente la borne supérieure pour les restrictions. Les listes L_{ASC} et L_{DESC} sont les suivantes :

L_{ASC}
SPERMIDINE : (POLYAMINE AMINE)
SPERMINE : (POLYAMINE AMINE) (AMINE_BIOGENE AMINE)
CHOLINE : (AMINE_BIOGENE AMINE)
CHOLINE : (AMINE_BIOGENE AMINE)

L_{DESC}
AMINE : (POLYAMINE AMINE_BIOGENE)
POLYAMINE : (SPERMINE, SPERMIDINE)
AMINE_BIOGENE : (SPERMINE, CHOLINE, HISTAMINE)

En pratique, *ctop* (ici AMINE) est retiré de L_{ASC} puisqu'il apparaît dans chaque entrée.

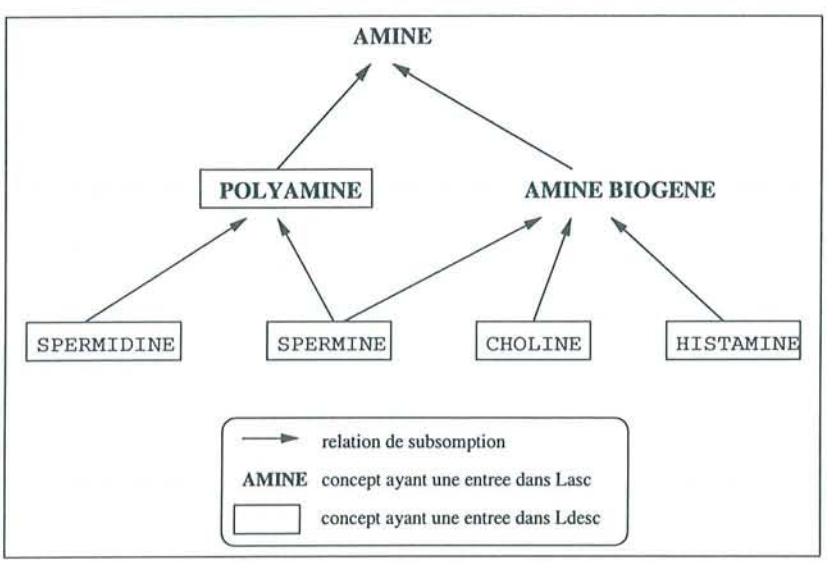

FIG. 4.8 – *Le graphe correspondant aux restrictions du rôle objet pour le sous-ensemble* SP_1

Le graphe construit à partir des restrictions et de leurs ascendants permet de regrouper les restrictions selon leur spécificités, et d'en déduire des généralisations par application de l'opération de ppsc sur les structures prédicatives associées aux restrictions.

Le parcours du graphe est effectué à partir de la borne supérieure des restrictions (algorithme 10). La fonction PARCOURIR-NOEUD (algorithme 11) traite les différents noeuds

du graphe.
Dans l'algorithme 11, les lignes 2 à 14 traitent le cas où le concept courant fait partie de la liste des restrictions. Dans ce cas, si le concept courant ne possède pas de descendant immédiat, il suffit de retourner la structure prédicative associée, ou le cas échéant la généralisation des structures prédicatives associées (fonction SP-ASSOCIEES).
S'il possède des descendants immédiats, alors chaque enfant est parcouru (ligne 7-9), et l'ensemble constitué du résultat de chaque enfant augmenté des structures prédicatives du concept courant (ligne 10) est généralisé (lignes 11-12).

Les lignes 15 à 32 considèrent le cas où le concept courant ne fait pas partie de la liste des restrictions. C'est donc un ancêtre d'une ou plusieurs restrictions.
S'il ne possède qu'un descendant immédiat, il suffit de traiter directement ce dernier, car le concept courant n'est qu'un concept intermédiaire sans intérêt (lignes 17-18).
Par contre, s'il possède plusieurs descendants immédiats, alors il constitue le PPSC d'un groupe de restrictions : chaque enfant est parcouru (lignes 21-23) et le résultat stocké dans la liste *Lvalide*.
L'ensemble constitué par *Lvalide* est ensuite généralisé (lignes 25-26), sauf si *Lvalide* est limité à un élément auquel cas on retrouve directement la structure prédicative correspondante (ligne 29). Ceci est rendu possible par le fait que deux enfants du concept courant peuvent retourner le même élément dans *Lvalide*.

Algorithme 10 Initialisation du parcours

DEBUT-PARCOURS(ctop)
- ctop : concept représentant la restriction maximale sur un rôle

LTOP = L_{DESC}(ctop)
Lvalide = nil
pour chaque c_i de LTOP **faire**
  Lvalide ← Lvalide ∪ PARCOURIR-NOEUD(c_i)
fin pour

L'algorithme 12 (fonction SP-ASSOCIEES) montre comment sont recupérées la ou les structures prédicatives associées à une restriction.
Dans le cas où une seule structure prédicative correspond à une restriction, elle est simplement retournée.
Dans le cas contraire, nous calculons la généralisation sur l'ensemble des structures prédicatives concernées et retournons cette généralisation.
Une alternative peut consister à appeler récursivement l'algorithme de parcours, afin de distinguer des généralisations plus fines. En effet, si plusieurs structures prédicatives possèdent la même restriction, elles sont généralisées ensemble, sans distinguer des ensembles plus fins. Toutefois, cela est en partie compensé par le traitement des autres rôles. D'autre part, nous préférons limiter le nombre de généralisation produites.

Dans notre exemple, pour le sous-ensemble de structures prédicatives SP_1 et le rôle *objet*, la fonction DEBUT-PARCOURS est appelée avec le paramètre ctop = AMINE. Nous avons donc (cf. figure 4.8) :

LTOP = {AMINE_BIOGENE, POLYAMINE}

Chaque élément de LTOP est examiné par la fonction PARCOURIR-NOEUD. Pour le

Algorithme 11 Traitement d'un noeud du graphe des restrictions

PARCOURIR-NOEUD(c)
- c : concept

1: **si** L_{ASC}(c) $\neq$ nil **alors**
2: **si** L_{DESC}(c) = nil **alors**
3: retourne SP-ASSOCIEES(c)
4: **sinon**
5: LC $\leftarrow$ L_{DESC}(c)
6: Lvalide $\leftarrow$ nil
7: **pour** chaque concept c_i de LC **faire**
8: Lvalide $\leftarrow$ Lvalide $\cup$ PARCOURIR-NOEUD(c_i)
9: **fin pour**
10: Lvalide $\leftarrow$ Lvalide $\cup$ SP-ASSOCIEE(c)
11: g $\leftarrow$ PPSC(Lvalide)
12: introduire g en CLASSIC et le marquer comme généralisation
13: retourne g
14: **fin si**
15: **sinon**
16: LC $\leftarrow$ L_{DESC}(c)
17: **si** TAILLE(L_{DESC}) = 1 **alors**
18: retourne PARCOURIR-NOEUD(PREMIER(LC))
19: **sinon**
20: Lvalide $\leftarrow$ nil
21: **pour** chaque concept c_i de LC **faire**
22: Lvalide $\leftarrow$ Lvalide $\cup$ PARCOURIR-NOEUD(c_i)
23: **fin pour**
24: **si** TAILLE(Lvalide) >1 **alors**
25: g $\leftarrow$ PPSC(Lvalide)
26: introduire g en CLASSIC et le marquer comme généralisation
27: retourne g
28: **sinon**
29: retourne SP-ASSOCIEES(PREMIER(Lvalide))
30: **fin si**
31: **fin si**
32: **fin si**

Algorithme 12 recherche des structures prédicatives associées à une restriction

SP-ASSOCIEES(c)
- c : concept représentant une restriction

LC ← Assoc(c)
si TAILLE(LC) = 1 **alors**
retourne PREMIER(LC)
sinon
g ← PPSC(LC)
introduire g en CLASSIC et le marquer comme généralisation
retourne g
fin si

concept AMINE_BIOGENE, nous somme dans le cas où le concept appartient à la fois à L_{ASC} et L_{DESC} car c'est une restriction qui possède des descendants (lignes 5-14).
Chacun des descendants est à son tour parcouru par PARCOURIR-NOEUD. Les trois concepts correspondants, CHOLINE, HISTAMINE et SPERMINE sont des restrictions sans descendants, et correspondent donc au cas où le concept appartient à L_{ASC} mais non à L_{DESC} (ligne 3).
L'appel à la fonction SP-ASSOCIEES retourne à chaque fois une structure prédicative initiale, puisque chaque restriction ne correspond qu'à une seule structure prédicative. Il en résulte que les valeurs retournées sont successivement : DOSAGE-3, DOSAGE-1 et DOSAGE-7. Nous avons donc :

$$Lvalide = \{\texttt{DOSAGE-3}, \texttt{DOSAGE-1}, \texttt{DOSAGE-7}, \texttt{DOSAGE-6}\},$$

suite à l'ajout de DOSAGE-6 (ligne 10).
La première généralisation calculée est le PPSC de *Lvalide* (lignes 11-12) :

```
G2-DOSAGE ≐ (and DOSAGE
                 (all objet AMINE_BIOGENE)
                 (all moyen CHROMATOGRAPHIE))
```

Ce traitement permet ainsi de généraliser les structures prédicatives ayant des restrictions de même niveau hiérarchique (CHOLINE, HISTAMINE et SPERMINE), ou d'un niveau hiérarchique immédiatement supérieur, puisque celui-ci est égal au PPSC de ces restrictions (ici AMINE_BIOGENE). Rappelons que ce traitement est celui du rôle *objet* : d'autres regroupements plus fins seront effectués pour le traitement du rôle *moyen*.

Il reste à parcourir le deuxième élément de LTOP, POLYAMINE. Nous sommes dans le cas où le concept appartient à L_{DESC} mais non à L_{ASC}. Le concept POLYAMINE possède deux descendant dans le graphe, le traitement est donc similaire à celui du concept AMINE_BIOGENE, à la différence près que seuls les descendants renvoient à des structures prédicatives, et qu'il n'y a donc aucune structure prédicative associée à POLYAMINE (cas des lignes 20-27). Nous avons donc :

$$Lvalide = \{\texttt{DOSAGE-4}, \texttt{DOSAGE-7}\}$$

La généralisation calculée est le PPSC de *Lvalide* (lignes 25-26) soit :

```
G3-DOSAGE ≐ (and DOSAGE
```

```
(all objet POLYAMINE)
(all moyen CHROMATOGRAPHIE))
```

Le parcours prend fin avec cette dernière généralisation pour le rôle *objet*.

Le même processus est appliqué sur le sous-ensemble SP_1 avec le rôle *moyen*. Le sous-graphe correspondant aux restrictions est reproduit figure 4.9. Nous avons alors *ctop* =

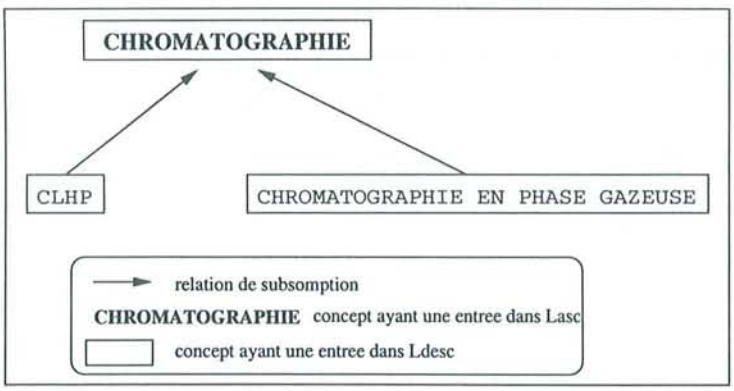

FIG. 4.9 – *Le graphe correspondant aux restrictions du rôle moyen pour le sous-ensemble* SP_1

CHROMATOGRAPHIE et

$$LTOP = \{\texttt{CLHP, CHROMATOGRAPHIE\_EN\_PHASE\_GAZEUSE}\}$$

L'application de PARCOURIR-NOEUD sur CLHP conduit à la généralisation des trois structures prédicatives associées à CLHP, {DOSAGE-1, DOSAGE-3, DOSAGE-7} (ligne 3) :

```
G1-DOSAGE ≐ (and DOSAGE
                 (all objet AMINE_BIOGENE)
                 (all moyen CLHP))
```

Le parcours de CHROMATOGRAPHIE_EN_PHASE_GAZEUSE retourne simplement la seule structure prédicative associée DOSAGE-6. Le parcours est alors fini. En effet, les valeurs retournées par le parcours des concepts de $LTOP$ ne sont pas exploitées car elles conduiraient à recalculer la généralisation sur l'ensemble SP_1.

Pour le sous-ensemble de structures prédicatives SP_2, il n'y a que deux structures prédicatives, et la seule généralisation possible a déjà été calculée (G2-DOSAGE). Aucun traitement n'est nécessaire.

Conclusion

L'ensemble des généralisations calculées sur les sept structures prédicatives DOSAGE-1, ..., DOSAGE-7 est résumé par le tableau 4.1 accompagné des structures prédicatives ayant contribuées à leur calcul. La hiérarchie obtenue après la généralisation est illustrée par la figure 4.10.

La première étape de la généralisation nous permet ainsi d'obtenir des structures prédicatives, regroupant les structures prédicatives initialement fournies, en se basant sur la

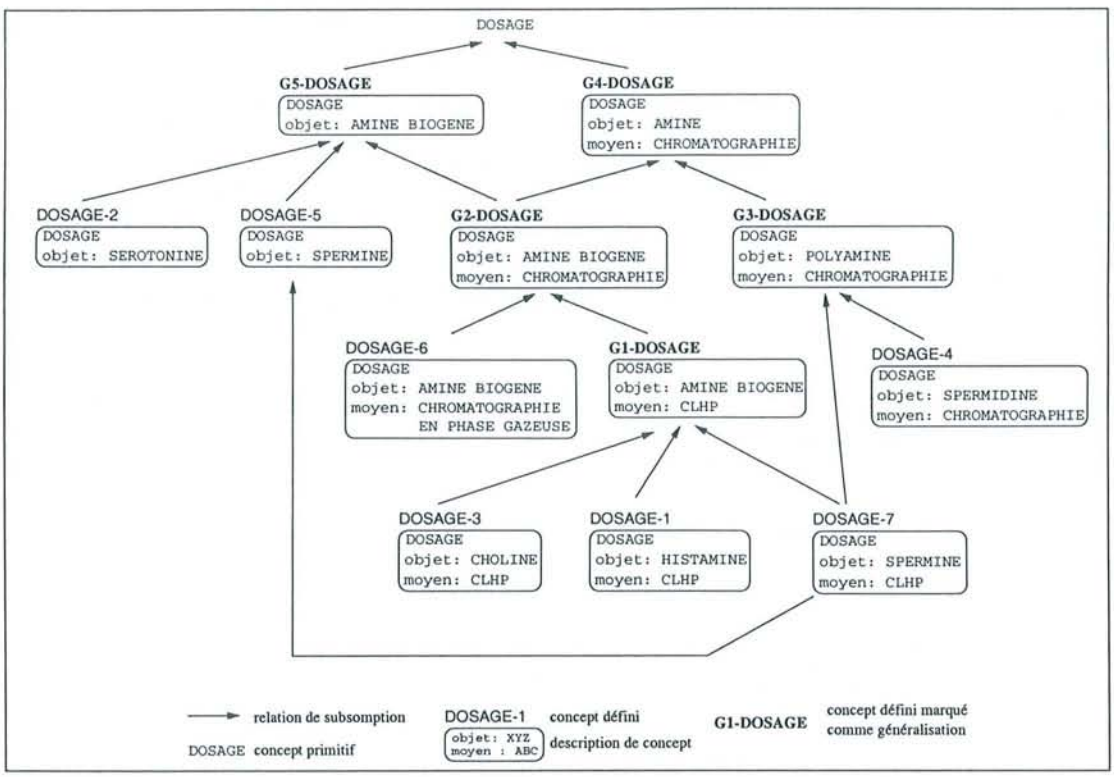

FIG. 4.10 – *Hiérarchie en* CLASSIC *après la première étape de généralisation*

généralité de leur arguments. Le concept `G1-DOSAGE` regroupe ainsi trois structures partageant le même *moyen*. Le concept `G2-DOSAGE` possède un niveau de généralité supérieur, et oppose la structure `DOSAGE-6` aux trois structures susdites par le *moyen* utilisé. La structure prédicative calculée la plus générale, `G4-DOSAGE`, constitue une bonne synthèse de l'information véhiculée par cinq structures (`DOSAGE-1, 3, 4, 6, 7`).
La présence de deux ensembles de rôles distincts, $\{objet, moyen\}$ et $\{objet\}$, contribue à la distinction de deux points de vue sur les données, l'un étant plus spécifique que l'autre car plus précis. On observera que le nombre de généralisation (cinq) est assez élevé par rapport au nombre de structures prédicatives initiales. Nous pensons donc que les choix effectués pour simplifier le processus, à différents endroits, sont justifiés : un traitement purement logique génère trop de données pour être exploitable.

Le traitement proposé repose toutefois sur des opérations logiques : la subsomption et le calcul du PPSC. La hiérarchie obtenue est facilement interprétable : la position de tel ou tel concept est indiscutable car elle découle de la subsomption entre les descriptions des concepts. Contrairement aux approches statistiques, on ne s'interroge pas sur les paramètres qui ont permis d'aboutir à telle ou telle hiérarchie de concepts. Le processus n'est cependant pas exhaustif, puisqu'un grand nombre de généralisations possibles ne sont pas effectuées : cela est indispensable pour ne pas obtenir un trop grand nombre de généralisations.

Cette première étape permet de traiter séparement les ensembles de structures prédicatives, prédicat par prédicat. La section suivante expose le processus qui permet de mettre en commun les structures prédicatives ayant des prédicats différents.

Généralisation	Structures prédicatives sources
G1-DOSAGE	DOSAGE-1, DOSAGE-3, DOSAGE-4, DOSAGE-6, DOSAGE-7
G2-DOSAGE	DOSAGE-1, DOSAGE-3, DOSAGE-6, DOSAGE-7
G3-DOSAGE	DOSAGE-4, DOSAGE-7
G4-DOSAGE	DOSAGE-1, DOSAGE-3, DOSAGE-7
G5-DOSAGE	DOSAGE-2, DOSAGE-5

TAB. 4.1 – *Généralisations et structures prédicatives ayant contribuées à les générer, pour les structures prédicatives* ***DOSAGE-1, ..., DOSAGE-7***

4.5 Deuxième étape de la généralisation : mise en commun des prédicats

Dans cette section, nous considérons des ensembles $PRED_i$ de structures prédicatives issus de la première étape de la généralisation, c'est-à-dire contenant déjà des généralisations, calculées selon la méthode présentée dans la section précédente. L'objectif est de calculer de nouvelles généralisations qui mettent en commun les informations relatives à plusieurs prédicats, en se basant une fois de plus sur l'opération de calcul de PPSC. Comme pour la première étape, nous présentons d'abord de manière informelle chaque étape du processus, suivi de l'algorithme correspondant, puis de son application sur un exemple. L'exemple considéré ici sera constitué de deux ensembles de structures prédicatives $PRED_1$ et $PRED_2$. L'ensemble $PRED_1$ comprend les structures prédicatives DOSAGE-1, DOSAGE-2, ..., DOSAGE-7, complétées par les généralisations calculées à la première étape. :

$PRED_1$ = {DOSAGE-1, DOSAGE-2, DOSAGE-3, DOSAGE-4, DOSAGE-5, DOSAGE-6, DOSAGE-7, G1-DOSAGE, G2-DOSAGE, G3-DOSAGE, G4-DOSAGE, G5-DOSAGE}

L'ensemble $PRED_2$ comprend un ensemble de structures prédicatives ayant la tête prédicative IDENTIFICATION et leurs généralisations issues de la première étape, dont la description textuelle en CLASSIC est la suivante :

```
IDENTIFICATION-1 ≐ (and IDENTIFICATION
                        (all objet SEROTONINE)
                        (all moyen CHROMATOGRAPHIE_SUR_GEL))
IDENTIFICATION-2 ≐ (and IDENTIFICATION
                        (all objet HISTAMINE)
                        (all moyen CHROMATOGRAPHIE_EN_PHASE_GAZEUSE))
IDENTIFICATION-3 ≐ (and IDENTIFICATION
                        (all objet SPERMIDINE)
                        (all moyen CLHP))
G1-IDENTIFICATION ≐ (and IDENTIFICATION
                         (all objet AMINE))
                         (all moyen CHROMATOGRAPHIE))
G2-IDENTIFICATION ≐ (and IDENTIFICATION
                         (all objet AMINE_BIOGENE)
                         (all moyen CHROMATOGRAPHIE))
```

La hiérarchie correspondant à $PRED_2$ est représentée sur la figure 4.11.

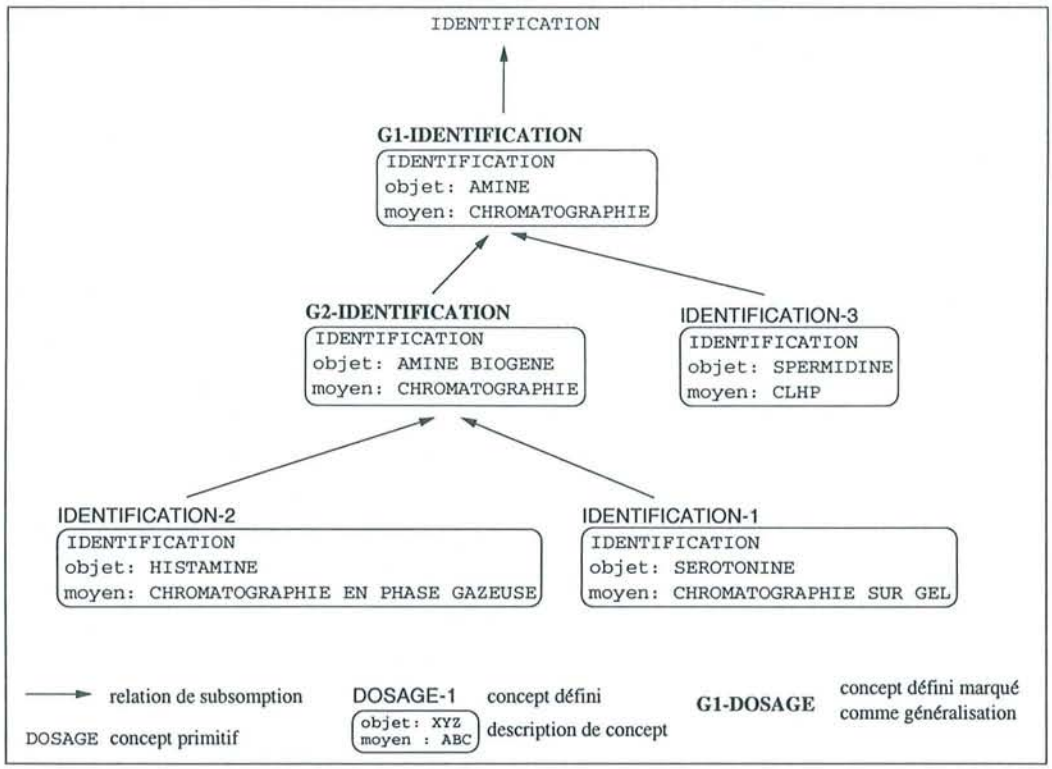

FIG. 4.11 – *Hiérarchie de concepts possédant la tête prédicative* `identification`

La deuxième étape de la généralisation repose sur les deux principes suivant :

- le regroupement des prédicats selon leur position hiérarchique, en utilisant la même méthode que celle utilisée pour les restrictions,
- la sélection des structures prédicatives les plus synthétiques dans chaque ensemble $PRED_i$, qui seules seront utilisées pour le calcul de nouvelles généralisations.

A ce stade, le regroupement des structures prédicatives repose donc sur les têtes prédicatives plutôt que sur les arguments associés aux relations. En effet, les arguments ont déjà été exploités lors de la première étape, et mis en commun à travers les généralisations calculées. Il s'agit maintenant de « remonter » les informations dans la hiérarchie, en regroupant prioritairement les structures prédicatives dont les prédicats sont proches dans la hiérarchie, de la même façon que cela a été fait pour les arguments.

Mais il ne s'agit pas de prendre en compte, à ce niveau, toutes les informations : c'est pourquoi nous choisissons de sélectionner, parmi chaque ensemble de structures prédicatives ayant une tête prédicative donnée, celles qui sont les plus générales. De cette façon ne sont mises en commun que des informations synthétiques, un nombre réduit de structures prédicatives qui représentent les informations principales véhiculées par l'ensemble des structures prédicatives extraites.

La première phase consiste à lister tous les prédicats qui apparaissent comme tête prédicative dans une des structures prédicatives initiales. A partir de cette liste, nous pouvons alors, comme nous l'avons fait pour les restrictions à la première étape, construire un graphe orienté à partir de la hiérarchie des concepts, et l'utiliser pour parcourir les

prédicats du plus spécifique au plus général. L'algorithme 13 (fonction GENERALISER-PREDICATS) correspond à cette première phase. La fonction PARCOURIR, qui permet de construire le graphe, est quasiment identique à celle utilisée dans la première étape (algorithme 9, page 69). La seule différence est l'appel de la fonction DEBUT-PARCOURS-2 à la place de la fonction DEBUT-PARCOURS. La différence avec la première étape se situe en effet, comme nous allons le voir plus loin, dans l'exploitation du graphe de concepts construit.

Algorithme 13 Traitement de l'ensemble des prédicats

GENERALISER-PREDICATS(b)
- b : booléen

LP = liste des concepts représentant les prédicats
PARCOURIR(TOP, LP) *;; TOP est l'élément maximum de la hiérarchie des concepts*

Dans notre exemple, les structures prédicatives sont limitées aux deux ensembles $PRED1$ et $PRED2$, la liste des prédicats est donc égale à :

$$LP = \{\texttt{DOSAGE}, \texttt{IDENTIFICATION}\}$$

Le sous-graphe correspondant est représenté figure 4.12. Lors de la deuxième étape, le paramètre *ctop* est toujours égal à la racine de la hiérarchie des concepts, le concept TOP, car toutes les structures prédicatives sont susceptibles d'être généralisées. En utilisant la hiérarchie de la figure 4.12, nous considérons *l'identification* comme une façon particulière de faire une *observation* et le *dosage* comme une façon particulière de faire une *analyse.*

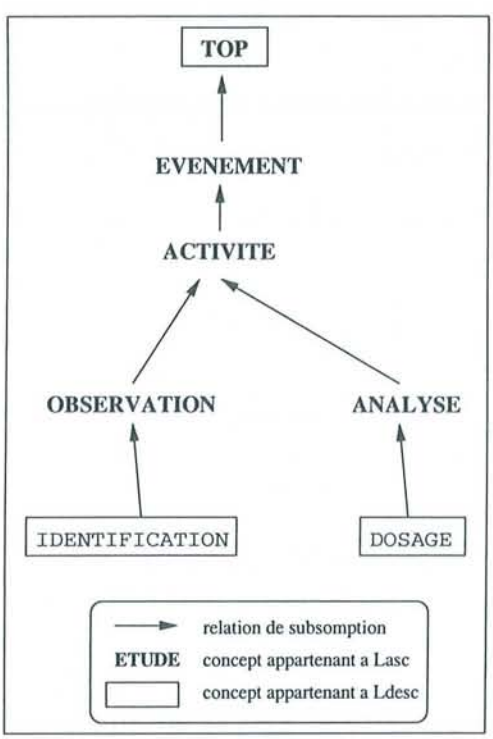

FIG. 4.12 – *Hiérarchie limitée aux prédicats et leurs ascendants*

Le parcours du graphe s'effectue à partir de la racine, de manière descendante. L'algorithme 14 (fonction DEBUT-PARCOURS-2) est de structure similaire à celui utilisé lors de la première étape. Une différence majeure est l'appel à la fonction GENERALISE-IPG (algorithme 16), utilisée pour sélectionner les structures prédicatives les plus générales pour calculer des généralisations (nous la décrivons page 82).
Dans notre exemple, nous avons $LTOP = \{\texttt{EVENEMENT}\}$.

Algorithme 14 canevas général du parcours

```
DEBUT-PARCOURS-2()
  LTOP = L_DESC(Ctop)
  Lvalide = nil
  pour chaque c_i de LTOP faire faire
    Lvalide ← Lvalide ∪ PARCOURIR-NOEUD-2(c_i)
  fin pour
  si TAILLE(Lvalide) >1 alors
    GENERALISE-IPG(Lvalide)
  fin si
```

Le parcours du graphe est détaillé par l'algorithme 15 (fonction PARCOURIR-NOEUD-2).
Les lignes 2 à 16 traitent le cas où le concept courant fait partie de la liste des prédicats. Dans ce cas, si le concept courant ne possède pas de descendant immédiat, il n'y a aucune généralisation à calculer, et le concept courant est retourné.
S'il possède des descendants, alors chaque enfant est parcouru (ligne 9-11), et le résultat retourné est ajouté à une liste.
Si cette liste n'est finalement pas limitée à un seul élément, une généralisation est calculée sur les structures prédicatives les plus générales associées aux prédicats de la liste, en appelant la fonction GENERALISE-IPG. On peut noter que le concept courant n'est pas ajouté à cette liste : seuls ses descendants immédiats sont considérés pour la généralisation. La prise en compte du concept courant n'est réalisée, le cas échéant, qu'au niveau de son ascendant direct. Cela contraste avec le traitement effectué lors de la première étape : il s'agit ici de « remonter » les informations à un niveau supérieur, sans intégrer directement ce dernier.

Les lignes 17 à 35 considèrent le cas où le concept courant ne fait pas partie de la liste des prédicats. C'est donc un parent d'un ou de plusieurs prédicats.
S'il ne possède qu'un descendant immédiat, il suffit de traiter directement ce dernier, car alors le concept courant n'est qu'un concept intermédiaire sans intérêt (lignes 19-21).
S'il possède plusieurs enfants, il correspond au PPSC d'un groupe de prédicats : chaque descendant est parcouru (lignes 25-27) et le résultat stocké dans la liste *Lvalide*. La fonction GENERALISE-IPG permet d'effectuer les généralisations correspondantes.

La fonction GENERALISE-IPG est donnée par l'algorithme 16. Cette fonction permet, à partir d'une liste de prédicats, de sélectionner les structures prédicatives les plus générales, et d'en calculer des généralisations. Pour cela, nous définissons pour un prédicat p,

Algorithme 15 Traitement d'un noeud du graphe des prédicats

```
PARCOURIR-NOEUD-2(c)
- c : concept
 1: si L_ASC(c) ≠ nil alors
 2:   si L_DESC(c) = nil alors
 3:     ;; cas où il n'y a plus de descendants
 4:     retourne c
 5:   sinon
 6:     ;; cas où il y a des descendants
 7:     LC ← L_DESC(c)
 8:     Lvalide ← nil
 9:     pour chaque concept c_i de LC faire faire
10:       Lvalide ← Lvalide ∪ PARCOURIR-NOEUD-2(c_i)      ;; si nouveau
11:     fin pour
12:     si taille(Lvalide) >1 alors
13:       GENERALISE-IPG(Lvalide)
14:       retourne c
15:     fin si
16:   fin si
17: sinon
18:   LC ← L_DESC(c)
19:   si taille(LC) = 1 alors
20:     ;; élément intermédiaire inutile
21:     retourne PARCOURIR-NOEUD-2(PREMIER(L_DESC))
22:   sinon
23:     ;; PPSC donc prise en compte
24:     Lvalide ← nil
25:     pour chaque concept c_i de LC faire faire
26:       Lvalide ← Lvalide ∪ PARCOURIR-NOEUD-2(c_i)      ;; si nouveau
27:     fin pour
28:     si TAILLE(Lvalide) >1 alors
29:       GENERALISE-IPG(Lvalide)
30:       retourne c
31:     sinon
32:       retourne PREMIER(Lvalide)
33:     fin si
34:   fin si
35: fin si
```

l'ensemble des informations les plus générales (IPG) :

> Étant donné un prédicat p, $IPG(p)$ est l'ensemble des structures prédicatives directement subsumées par p.

Ce choix pour sélectionner les structures prédicatives repose sur plusieurs considérations :

- les structures prédicatives les plus générales sont par définition celles qui synthétisent le plus d'informations,
- la sélection est très simple à réaliser,
- le nombre de structures prédicatives sélectionnées est réduit.

Une alternative possible consiste à définir un critère plus fin de pertinence et de généralité pour les structures prédicatives à sélectionner. La figure 4.13 illustre par exemple un cas où les structures sont inégalement distribuées : le concept `SP-2` est qualitativement beaucoup plus important que le concept `SP-3` puisqu'il synthétise toutes les structures prédicatives correspondant au prédicat p sauf une. Dans ce cas, nous avons

$$IPG(p) = \{ \texttt{SP-1} \}$$

La sélection de la structure `SP-2` peut cependant s'avérer plus pertinente que celle de `SP-1`, par exemple, s'il existe une différence importante de généralité entre les deux concepts. En effet, une structure prédicative trop générale fournit peu d'information, et dans ce cas `SP-2` « couvre » une majeure partie de l'information, puisqu'elle subsume toutes les structures à l'exception de `SP-3`.
Il peut alors être tentant de définir une heuristique qui permettent de sélectionner des structures moins générales et plus informatives. Toutefois, nous n'avons pas retenu cette solution car elle nécessite d'utiliser des critères numériques pour estimer la couverture d'une structure prédicative (seuil, pourcentage, ...), qui sont généralement artificiels et difficiles à établir.

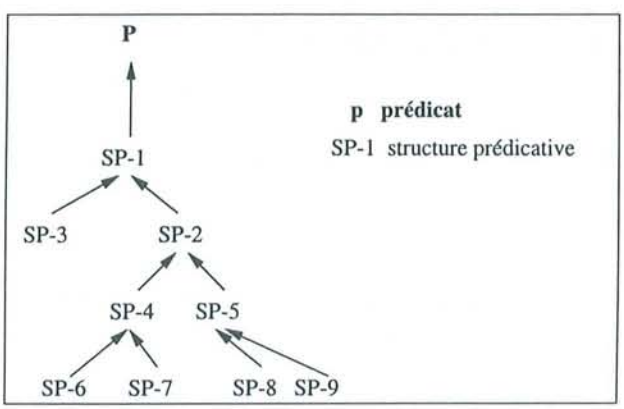

FIG. 4.13 – *Un exemple de distribution non uniforme de structures prédicatives*

Après la sélection des structures prédicatives pour deux ou plusieurs prédicats, celles-ci sont triées selon leur ensemble de rôles afin de respecter l'heuristique $\mathcal{H}_2$ (lignes 4-7 de l'algorithme 16). La généralisation est réalisée directement sur chaque ensemble de structures prédicatives ainsi obtenues par application de l'opérateur `ppsc` (lignes 9-15). Il est possible d'adopter un traitement plus complet, en appliquant sur ces ensembles le processus de

l'étape numéro un, c'est-à-dire en tenant compte de la spécificité des arguments des différentes structures prédicatives. Toutefois, l'objectif de cette étape est d'obtenir un nombre réduit de structures prédicatives assez générales, s'appuyant sur celles déjà calculées lors de la première étape. Appliquer un processus trop fin conduirait ici à générer des structures prédicatives peu intéressantes, ajoutant des niveaux hiérarchiques supplémentaires inutiles dans la hiérarchie des concepts. Nous nous limitons donc au traitement basique indiqué.

Algorithme 16 Généralisation des informations les plus générales d'une liste de prédicats

GENERALISE-IPG(LP)
- LP : liste de concepts représentant des prédicats

1: TableSP ← nil ;; *table contenant les struct. préd. indexées par liste de rôles*
2: **pour** chaque prédicat p de LP faire **faire**
3: IPG_p ← ensemble des structures prédicatives directement subsumées par p
4: **pour** chaque concept c de l'ensemble IPG_p faire **faire**
5: index ← LISTE-ROLES(c)
6: TableSP(index) ← TableSP(index) ∪ c
7: **fin pour**
8: **fin pour**
9: **pour** chaque index de TableSP faire **faire**
10: LC ← TableSP(index)
11: **si** TAILLE(LC) >1 **alors**
12: g ← PPSC(LC)
13: introduire g en CLASSIC et le marquer comme généralisation
14: **fin si**
15: **fin pour**

L'application du processus de parcours à notre exemple donne le résultat suivant : les concepts `EVENEMENT` et `ACTIVITE` sont parcourus sans aucun effet, puisqu'ils n'appartiennent pas à la liste des prédicats et ne possèdent qu'un enfant (lignes 3-4). Le concept `ACTIVITE` possède deux descendants : `OBSERVATION` et `ANALYSE`, qui sont tous deux des concepts intermédiaires absents de la liste des prédicats, et dont le parcours renvoie respectivement les concepts `IDENTIFICATION` et `DOSAGE`.
Nous avons donc $Lvalide$ = {`IDENTIFICATION`, `DOSAGE`}.

La fonction GENERALISE-IPG s'applique sur $Lvalide$. On calcule pour chaque prédicat p son ensemble $IPG(p)$:

IPG(`IDENTIFICATION`) = {`G1-IDENTIFICATION`}
IPG(`DOSAGE`) = {`G4-DOSAGE`, `G5-DOSAGE`}

La table $TableSP$ est donnée par :

$TableSP$({objet, moyen}) = {`G4-DOSAGE`, `G1-IDENTIFICATION`}
$TableSP$({objet}) = {`G5-DOSAGE`}

Seul le premier index conduit à une généralisation, le deuxième étant réduit à un élément. Cette généralisation est la suivante :

```
PPSC({G4-DOSAGE, G1-IDENTIFICATION}) →
    G1-ACTIVITE ≐ (and ACTIVITE
                       (all objet AMINE)
                       (all moyen CHROMATOGRAPHIE))
```

Cette généralisation est introduite dans la hiérarchie de concepts, et le processus prend fin.

Conclusion

Nous illustrons le résultat de cette deuxième étape du processus à l'aide de trois schémas : la figure 4.14 montre tout d'abord la hiérarchie conceptuelle telle qu'elle se présente à la fin de la première étape du processus de généralisation. La figure 4.15 montre la hiérarchie conceptuelle obtenue à la fin de la deuxième étape du processus. Enfin la figure 4.16 montre une vue partielle de la hiérarchie finale, où seules apparaissent les structures prédicatives, initiales et calculées. Sur chacune des ces figures, seules les structures prédicatives les plus générales apparaissent, pour ne pas alourdir la présentation.

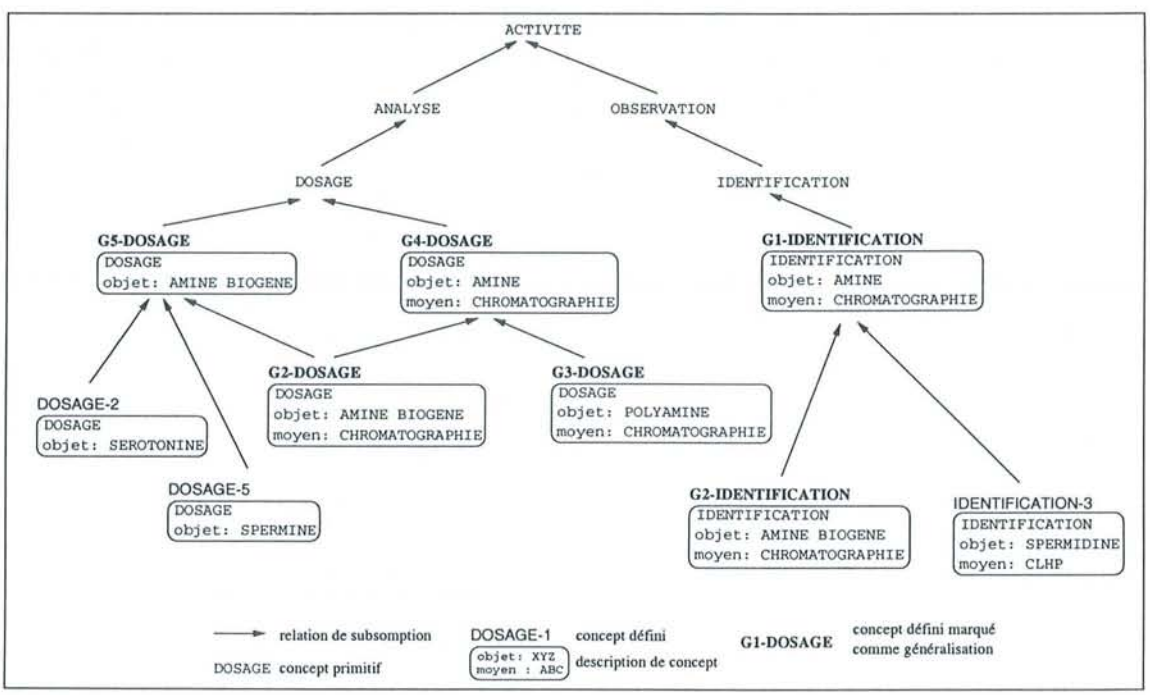

FIG. 4.14 – *Hiérarchie des concepts avant la deuxième étape de la généralisation*

La première figure (4.14) met en évidence les généralisations issues de la première étape, limitées à un seul prédicat. La deuxième figure (4.15) montre les nouveaux liens mis en place après l'introduction de la généralisation `G1-ACTIVITE`. Enfin, la figure 4.16 permet une meilleure appréciation de la hiérarchie : c'est celle qui est montrée à l'utilisateur (sans le trait de séparation), car celui-ci ne s'intéresse *a priori* qu'aux structures prédicatives. Ainsi sont mis en évidences deux structures prédicatives `G5-DOSAGE` et `G1-ACTIVITE` qui couvrent

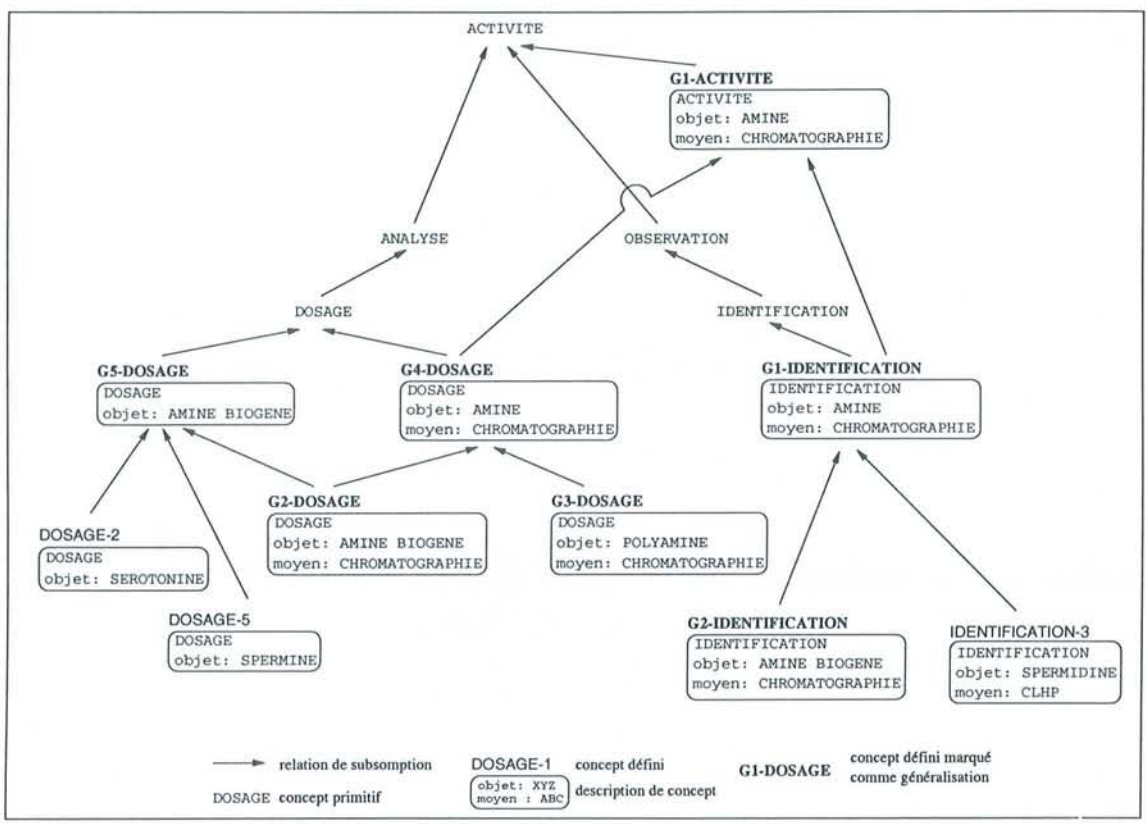

FIG. 4.15 – *Hiérarchie des concepts après la deuxième étape de la généralisation*

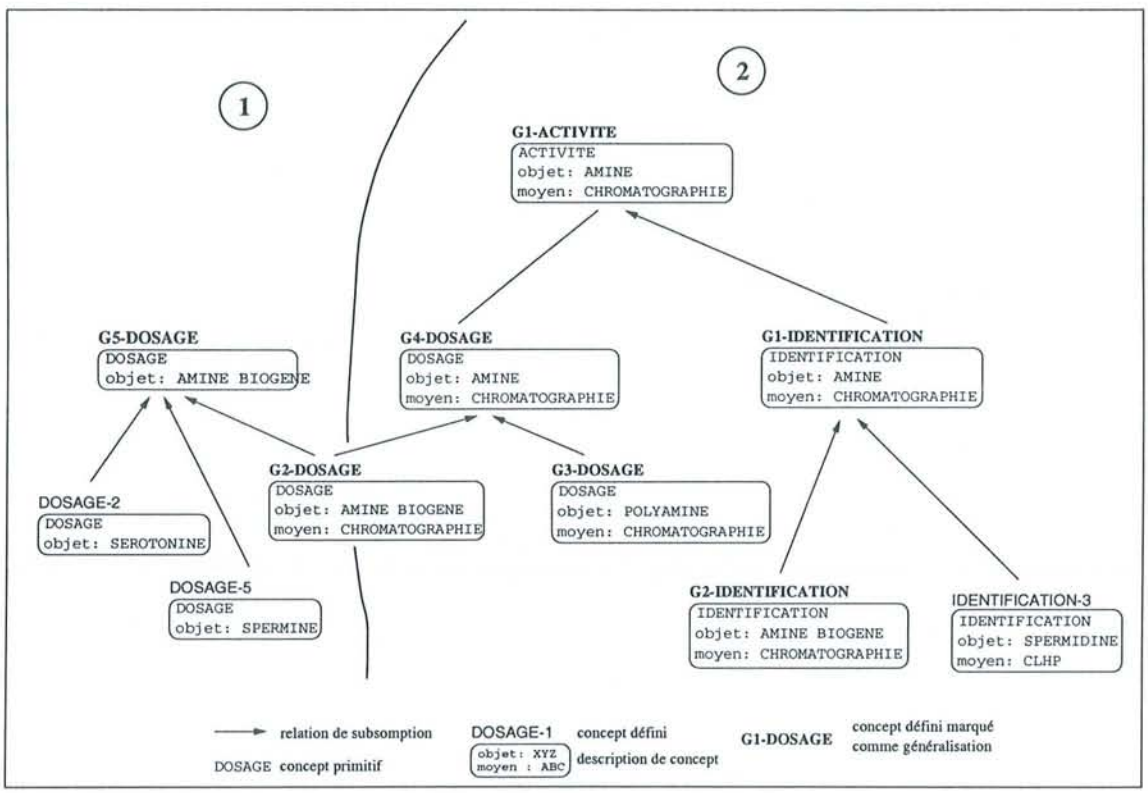

FIG. 4.16 – *Hiérarchie des concepts où seules apparaissent les structures prédicatives*

ensemble la totalité des dix structures prédicatives initiales. La généralisation `G1-ACTIVITE` offre une synthèse pertinente, compréhensible, d'un ensemble de huit structures prédicatives initiales. La généralisation `G5-DOSAGE` est moins importante quantitativement, mais offre une information plus précise au niveau des termes employés. Cette double couverture est schématisée sur la figure par la séparation en deux parties des différentes structures prédicatives. Le processus de généralisation permet ainsi de synthétiser les informations extraites des textes en offrant un nombre réduit de structures prédicatives plus générales.

Nous nous intéressons dans la section suivante à la complexité calculatoire du processus de généralisation, avant de conclure sur les apports de notre méthode (section 4.7).

4.6 Complexité du processus de généralisation

Un processus de généralisation est, par nature, gourmand en temps de calcul. Il est donc nécessaire de s'assurer que la complexité calculatoire du processus n'est pas un obstacle insurmontable à son utilisation. Nous donnons dans un premier temps une estimation de la complexité théorique en temps de notre processus (section 4.6.1). Nous la complétons par une évaluation empirique sur des données de tailles variables afin d'estimer plus pragmatiquement le temps nécessaire pour une utilisation réelle du processus (section 4.6.2).

4.6.1 Complexité théorique

L'opération de base utilisée par les algorithmes est l'opération de PPSC. Sa complexité est en $\Theta(N)$ pour le traitement d'un ensemble de N concepts. En effet, soit n_r le nombre de rôles maximum d'une structure prédicative, n_n le nombre maximal de structures prédicatives apparaissant dans une autre structure[36], $Prof$ la profondeur maximale de la hiérarchie, l'algorithme procède au maximum à $(n_r + 1) \times n_n \times Prof$ opérations pour calculer la généralisation de deux concepts. Nous supposons ici que le test de subsomption fourni par CLASSIC prend un temps constant, ce qui est vérifié pour des concepts peu complexes, et vérifié en pratique. Les trois paramètres n_n, n_r, $Prof$ peuvent être bornés par un nombre fixe. n_n et n_r sont toujours inférieurs à 10 en pratique. $Prof$ est en fait borné car il dépend de la hiérarchie des termes primitifs qui est fixée : seul les concepts structures prédicatives sont ajoutés à la hiérarchie. Pour N concepts, la complexité est de $(n_r + 1) \times n_n \times Prof \times N$, soit $\Theta(N)$. Cette opération est donc peu coûteuse. C'est son application sur de grand ensemble de structures prédicatives qui est problèmatique.

La première étape de la généralisation est de complexité plus élevée. Le traitement de N structures prédicatives ayant le même ensemble de rôles s'applique sur n_r rôles. La liste des restrictions et son inverse se calculent en $\Theta(N)$. Dans le cas où la hiérarchie est stricte, la parcours du graphe des restrictions effectue au maximum $N \times Prof$ calculs de PPSC, ce qui donne une complexité en $\Theta(N)$. Cependant, la hiérarchie utilisée est a priori non stricte. Le coût du parcours du graphe peut alors être nettement plus élevé : une structure prédicative peut être utilisée plusieurs fois pour le calcul du PPSC. La complexité dépend alors fortement du taux de branchement (le nombre de descendants immédiats d'un noeud)

36. Lorsqu'une structure est imbriquée dans une autre. Ce cas est illustré par les concepts `SIMULATION-1` et `SIMULATION-2` de la section 4.1, où $n_n = 1$. En pratique n_n est toujours un petit nombre, et souvent égal à zéro.

et du nombre moyen de parents d'un noeud. Un calcul fin de la complexité dans ce cas est difficile. Pour les cas vraiment défavorables, la complexité peut être en $\Theta(N^2)$.

La deuxième étape est similaire à la première en terme de complexité, puisque les algorithmes utilisés sont structurellement équivalents. La complexité se trouve donc située selon les structures hiérarchiques utilisées entre $\Theta(N)$ et $\Theta(N^2)$.

La complexité de notre processus est, en toute logique, comparable à celle de la méthode COING présentée section 2.5.3. Elle n'est toutefois qu'indicative et dépend de divers paramètres : nous avons déjà remarqué l'importance de la structure hiérarchique sur la complexité du processus ; à cela s'ajoute le fait que notre calcul est également dépendant du nombre de prédicats utilisés et du nombre de structures prédicatives possédant un prédicat donné. Nous proposons donc d'effectuer une évaluation empirique de la complexité de notre processus. Celle-ci est présentée dans la section suivante.

4.6.2 Evaluation empirique de la complexité

L'évaluation empirique que nous avons effectuée prend en compte les divers paramètres qui peuvent influencer la complexité du processus. Pour construire une base de structures prédicatives, nous avons utilisé une hiérarchie de termes dérivée d'un thésaurus utilisé pour notre application. Les détails sur ce thésaurus sont donnés dans le chapitre décrivant notre application (chapitre 6). Il suffit à ce stade de savoir que l'on dispose d'un ensemble de 8000 termes environs, constituant une hiérarchie de profondeur maximale $P = 20$. Parmi ces termes, 450 environ sont utilisés comme prédicats. La hiérarchie n'est pas stricte, et reflète une utilisation réelle puisque nous l'avons utilisée pour notre expérimentation (i.e, ce n'est pas une hiérarchie artificielle). On compte 610 termes possédant deux parents et 30 termes possédant trois parents.

A partir de cette hiérarchie, nous avons construit des bases de structures prédicatives de tailles différentes. Dans toutes les bases, les structures prédicatives sont construites de manière aléatoire, en prenant un prédicat parmi les 450 disponibles et des arguments parmi les 8000 termes. Lorsque nous faisons des tests avec des prédicats différents, nous vérifions que le même prédicat n'est pas choisi deux fois. Les structures prédicatives obtenues n'ont bien entendu aucune signification. Elles sont a priori beaucoup plus hétérogènes que dans un cas réel, puisque les arguments d'un même prédicat peuvent être très éloignés dans la hiérarchie. Les tests constituent donc des cas défavorables par rapport à une utilisation « normale ».

Les paramètres que nous avons pris en compte pour cette évaluation sont les suivants :

- le nombre de structures prédicatives ayant une tête prédicative donnée, N_{sp}. Par exemple, s'il y a trois structures prédicatives possédant le prédicat P_1, trois structures possédant le prédicat P_2, et ainsi de suite, alors $N_{sp} = 3$,
- le nombre de têtes prédicatives différentes, N_{tp},
- le nombre de rôles que possède chaque structure prédicative, N_r.

Les deux premiers paramètres donnent le nombre total de structures prédicatives traitées : $N_{total} = N_{sp} \times N_{tp}$.

Le comportement du processus de généralisation étant très lié au nombre N_{sp}, nous avons tout d'abord testé des bases de structures prédicatives avec un seul prédicat, soit

$N_{tp} = 1$, en faisant varier la valeur de N_{sp}. Deux valeurs différentes de N_r ont été utilisées, respectivement $N_r = 2$ et $N_r = 4$. Nous avons fait varier N_{sp} pour des valeurs comprises entre 5 et 2000. Les résultats sont donnés par le graphe 4.17 pour $N_r = 2$ et $N_r = 4$. Ce

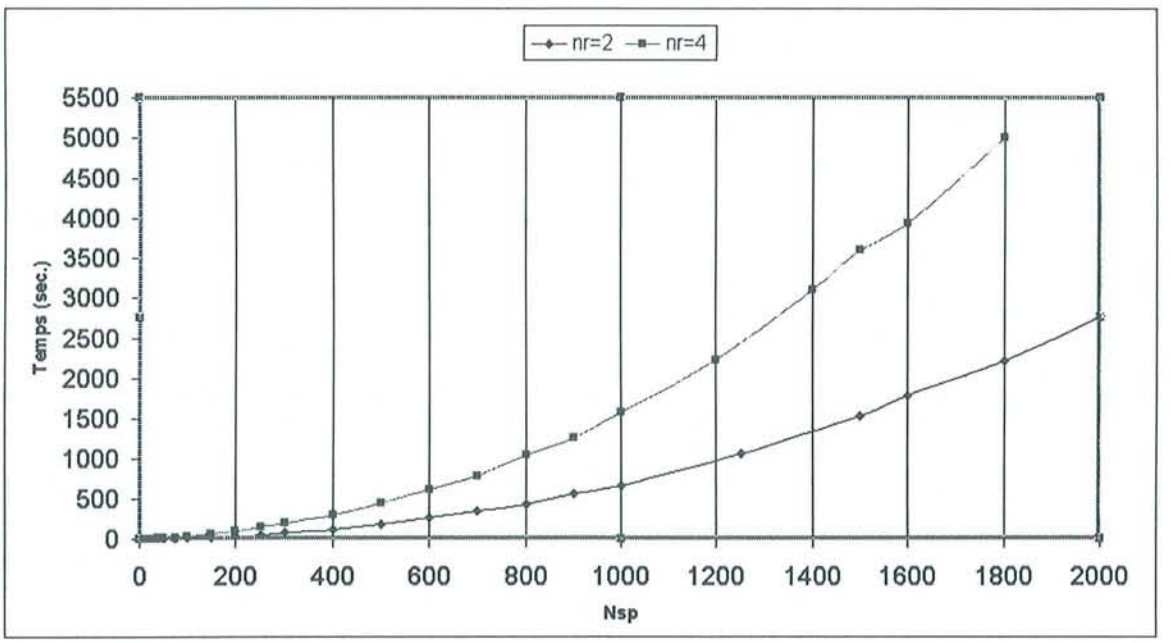

FIG. 4.17 – *Temps de généralisation en secondes en fonction de N_{sp} pour $N_{tp} = 1$ et deux valeurs de N_r : 2 et 4*

premier test (figure 4.17) permet de confirmer les valeurs théoriques de complexité établies dans la section précédente : le temps de calcul n'est pas linéaire en fonction de N_{sp}, mais croît plus lentement que N_{sp}^2.
Ce résultat est comparable, voire légérement supérieur, à celui obtenu par Bournaud [Bournaud 96, p. 136] : le système COING permet par exemple de calculer les généralisations de 400 objets, possédant en moyenne 9 relations, en 14 minutes. Dans le cas de la figure 4.17, notre processus traite environ 1100 structures prédicatives pour $N_r = 2$, et 600 structures pour $N_r = 4$, en 14 minutes.
Ces chiffres ne sont bien sûr qu'indicatifs : les hiérarchies utilisées sont différentes (Bournaud utilise un treillis de types de profondeur maximale égale à 4), et les structures prédicatives sont différentes des objets de Bournaud. Le système COING semble toutefois plus limité en mémoire, puisque Bournaud n'a pu l'exécuter sur une base de 900 objets (dépassement de mémoire de l'interpréteur Lisp utilisé). Pour $N_r = 4$, la limite de notre processus a été atteinte pour 2000 structures prédicatives (pour la même raison).

Le deuxième point important à tester est le comportement du processus avec un nombre important de structures prédicatives, réparties entre plusieurs prédicats. Dans ce cas, nous avons fait varier N_{sp} en proportions moins grandes, entre 5 et 50. Le nombre N_{tp} varie quant à lui entre 5 et 200. Nous avons utilisé deux valeurs différentes de N_r, respectivement $N_r = 2$ et $N_r = 4$. Les résultats sont illustrés par les graphes 4.18 et 4.19 pour des structures prédicatives possédants 2 rôles et par les graphes 4.20 et 4.21 pour des structures prédicatives possédants 4 rôles.
Sur la figure 4.18, le temps de calcul est représenté en fonction du nombre de prédicats

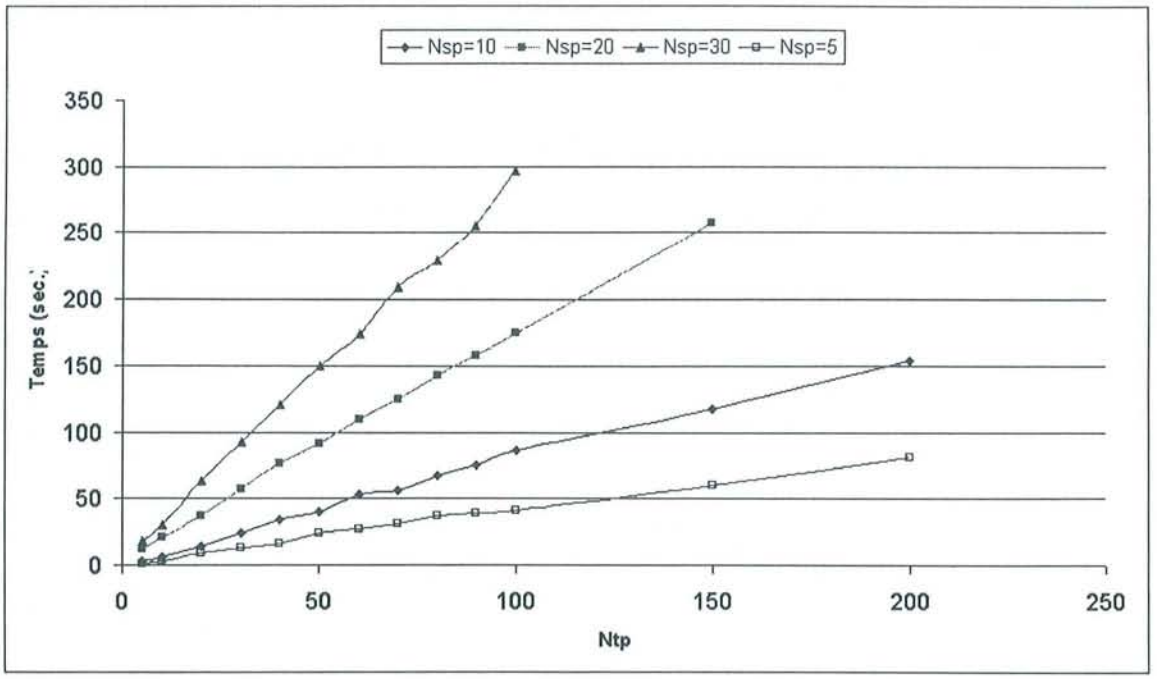

FIG. 4.18 – *Temps de généralisation en secondes en fonction de N_{tp} pour des valeurs de N_{sp} comprises entre 5 et 30, et $N_r = 2$*

différents utilisés (N_{tp}). Pour les quatre valeurs différentes de N_{sp} utilisées, la progression est linéaire : ceci semble montrer que le principal coût de l'algorithme est fonction de N_{sp} et non du nombre de prédicats utilisés.

La figure 4.19 représente le temps de calcul en fonction du nombre total de structures prédicatives (N_{total}). Nous constatons que pour un nombre N_{total} donné, les temps de calcul pour différentes valeurs de N_{sp} sont similaires. Ils ont cependant tendance à augmenter avec une valeur de N_{sp} croissante.

Les temps de calculs sont sensiblement moins élevés que dans le premier test où une valeur de $N_{tp} = 1$ est utilisée. Cela s'explique par le fait que les valeurs de N_{sp} sont beaucoup moins grandes. Ce résultat est très intéressant puisque nous sommes dans un cas plus « plausible » que pour le premier test. Ainsi, notre processus traite 3000 structures prédicatives en moins de 300 secondes pour des valeurs de N_{sp} égales à 20 et 30.[37]

Les figures 4.20 et 4.21 appellent globalement les mêmes commentaires que les deux précédentes, seule la valeur de N_r étant modifiée ($N_r = 4$). Nous constatons toutefois que les temps de calcul sont légérement supérieurs au double des valeurs constatées pour $N_r = 2$, ce qui indiquerait une progression non linéraire. Par exemple, il faut environ 150 secondes pour traiter 1500 structures prédicatives avec $N_{sp} = 30$ et $N_r = 2$ contre un peu moins de 350 secondes pour traiter 1500 structures prédicatives avec $N_{sp} = 30$ et $N_r = 4$. Cependant la valeur $N_r = 4$ représente un maximum possible dans un cas d'utilisation réel, et les temps calculés restent « corrects ». La capacité mémoire semble être atteinte autour de 2000 structures prédicatives traitées.

37. Nous nous sommes limités à $N_{tp} = 200$, ce qui explique que le chiffre de 3000 structures n'est pas

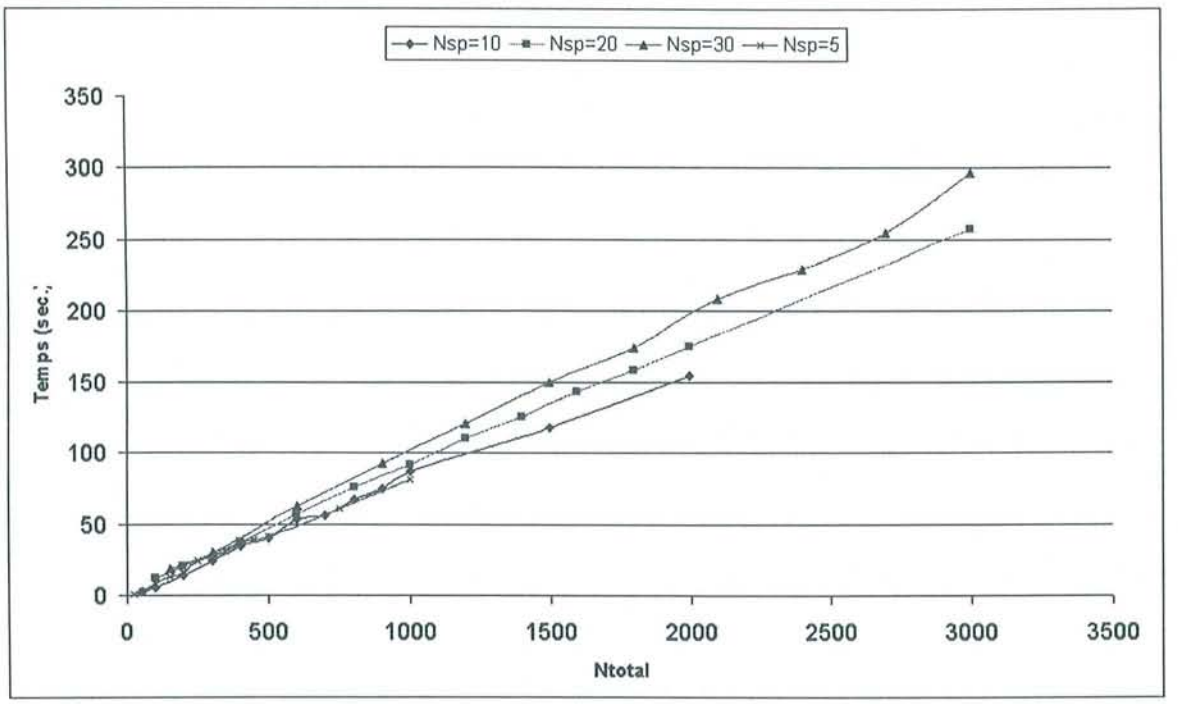

FIG. 4.19 – *Temps de généralisation en secondes en fonction de N_{total} pour des valeurs de N_{sp} comprises entre 5 et 30, et $N_r = 2$*

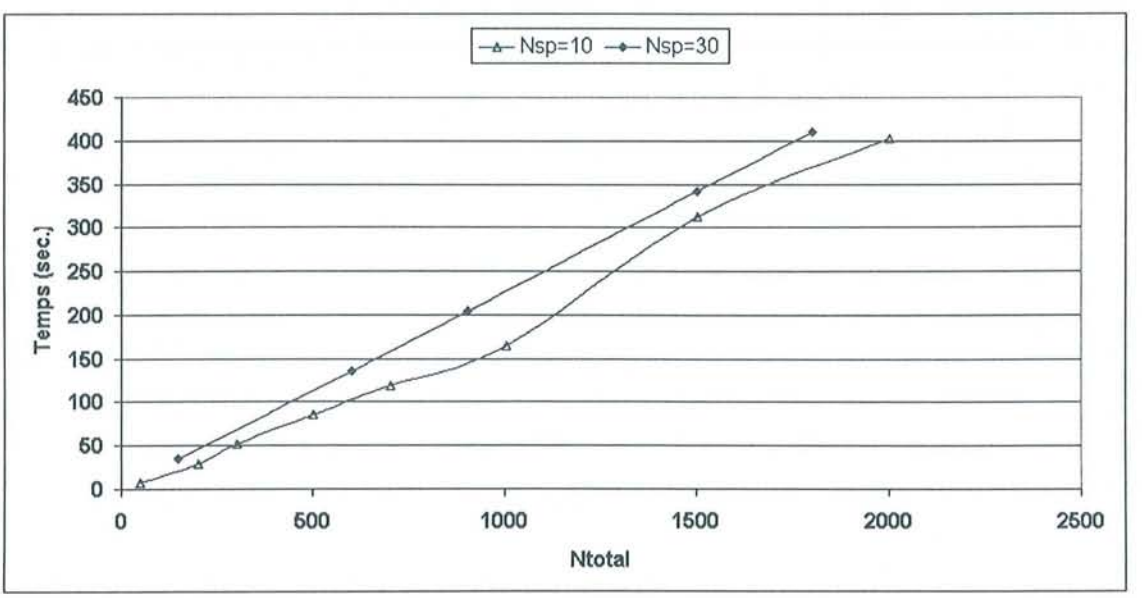

FIG. 4.20 – *Temps de généralisation en secondes en fonction de N_{tp} pour des valeurs de N_{sp} comprises entre 10 et 30 et $N_r = 4$*

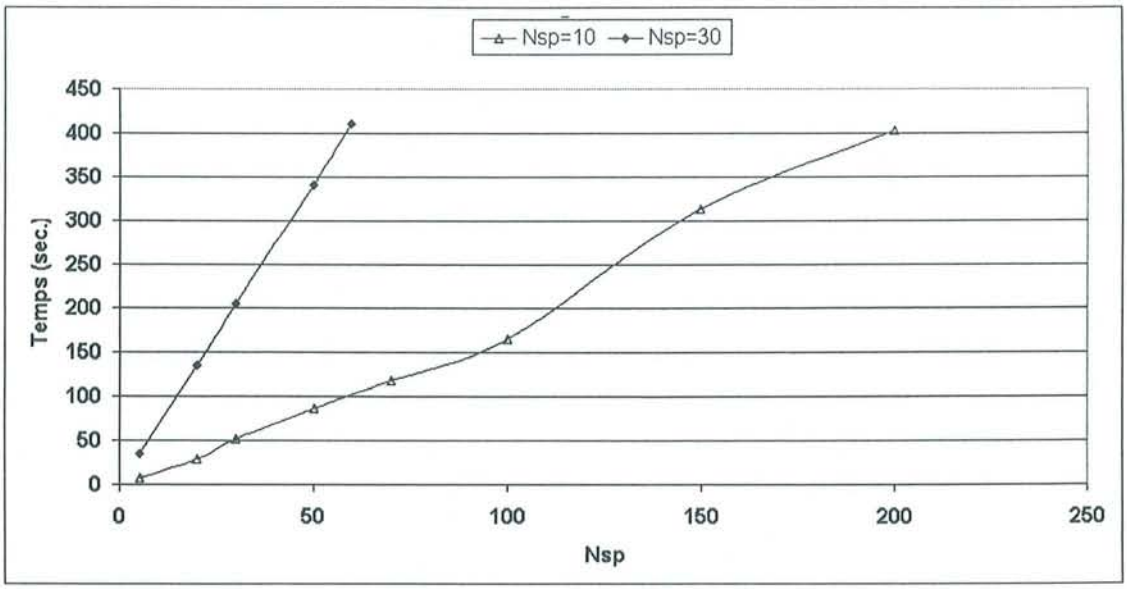

FIG. 4.21 – *Temps de généralisation en secondes en fonction de N_{tp} pour des valeurs de N_{sp} comprises entre 10 et 30 et $N_r = 4$*

Notre dernier test tente de reproduire de façon plus exacte une situation réelle. Les valeurs de N_{tp} et de N_{sp} utilisées ne sont pas uniformes pour la base de structures prédicatives, mais échelonnées de manière régulière. Dans les deux exemples ci-dessous, quelques prédicats possèdent un grand nombre de structures prédicatives associées, tandis que la majorité des prédicats ne possèdent que quelques structures.

Exemple 1 :
1 prédicat avec $N_{sp} = 50$, i.e. 50 structures prédicatives possèdent ce prédicat,
2 prédicats avec $N_{sp} = 40$,
3 prédicats avec $N_{sp} = 30$,
4 prédicats avec $N_{sp} = 20$,
5 prédicats avec $N_{sp} = 10$,
50 prédicats avec $N_{sp} = 5$.

Pour un total de 600 structures prédicatives réparties sur 65 prédicats et possédant 2 rôles ($N_r = 2$), le temps de calcul est de 68 secondes, et conduit à 678 généralisations.

Exemple 2 :
1 prédicat avec $N_{sp} = 150$,
2 prédicats avec $N_{sp} = 100$,
3 prédicats avec $N_{sp} = 75$,
4 prédicats avec $N_{sp} = 50$,
5 prédicats avec $N_{sp} = 30$,
10 prédicats avec $N_{sp} = 20$,
20 prédicats avec $N_{sp} = 10$,
50 prédicats avec $N_{sp} = 5$,

atteint pour les valeurs 5 et 10 de N_{sp}.

100 prédicats avec $N_{sp} = 3$,
250 prédicats avec $N_{sp} = 2$.

Pour un total de 2375 structures prédicatives réparties sur 445 prédicats et possédant 2 rôles ($N_r = 2$), le temps de calcul est de 232 secondes (soit un peu moins de 4 minutes), et conduit à 2221 généralisations. Cela confirme que le temps de calcul dépend surtout de N_{sp}, qui, en pratique, n'est élevé que pour un nombre réduit de prédicats. Le nombre élevé de généralisations calculées n'est pas étonnant dans la mesure où les structures prédicatives utilisées sont très hétérogènes.

Cette évaluation empirique montre que notre processus peut être appliqué à plusieurs milliers de structures prédicatives extraites de textes, dans un temps que l'on peut estimer raisonnable pour une utilisation réelle.

4.7 Conclusion

Nous avons présenté un processus de généralisation de structures prédicatives scindé en deux étapes, en nous basant sur un ensemble de principes et d'heuristiques pour limiter la complexité calculatoire et le nombre de généralisations produites. Nous concluons en mettant en évidence les spécificités de notre solution par rapport à celles utilisées en classification conceptuelle et présentés dans le chapitre 2.

Tout d'abord, notre processus se distingue nettement des approches classiques (section 2.3) : l'algorithme ne repose pas sur une fonction d'évaluation et ne nécessite pas de fixer des paramètres de fonctionnement. Le regroupement des structures prédicatives (équivalent des objets utilisés pour former des classes) est basée sur une opération définie logiquement, et la construction des généralisations (équivalent des classes) peut être facilement appréhendée. Les approches classiques fonctionnent au contraire comme des boites noires qui fournissent une hiérarchie dont la construction est cachée. De plus, pour les systèmes incrémentaux, la hiérarchie résultante dépend de l'ordre d'introduction. Notre solution est non incrémentale et unique pour un ensemble de structures prédicatives données. La position relative d'une structure prédicative dans la hiérarchie découle uniquement de propriétés logiques (la subsomption), alors que dans un système classique, la position d'un objet dépend du processus de construction et de la fonction d'évaluation associée. Enfin nous utilisons des connaissances de domaine (la hiérarchie sur les termes utilisés dans les structures prédicatives), ce qui est rarement le cas dans les approches classiques.

Notre processus de généralisation partage avec les solutions utilisant les treillis de concepts, la représentation par objet et les graphes conceptuels une approche logiquement fondée, utilisant des connaissances du domaine. Les connaissances du domaine consistent en un ensemble de concepts organisés en hiérarchie. Dans tous ces travaux, la hiérarchie obtenue est unique. Cependant, ces approches sont caractérisées par la recherche de toutes les classes possibles à partir d'un ensemble d'objets. Cela signifie qu'il n'y a pas d'heuristiques pour réduire le nombre de classes générées, contrairement à la solution que nous avons présentée. Une exception est le travail sur la MSG et l'amélioration proposée par Bournaud : les graphes conceptuels sont décomposés pour obtenir des arcs et réduire ainsi le nombre de généralisations possibles. Cependant, à partir des arcs, toutes les généralisations possibles sont calculées, et le processus proposé par les auteurs ne procède donc à aucun élagage.

En ce qui concerne les travaux de Carpineto et Romano [Carpineto 96], les objets considérés sont des descripteurs de document : ce sont donc des objets peu complexes, et les hiérarchies obtenues peuvent être interprétées même en générant toutes les classes possibles. Appliquer ce même principe à des structures prédicatives aboutirait à des hiérarchies très complexes. Nous sommes donc obligés de proposer des heuristiques qui permettent de réduire les généralisations calculées. En cela, notre processus est moins exhaustif, mais plus approprié au traitement des structures prédicatives.

Les travaux de Simon et Napoli [Simon 98], de Mineau [Mineau 90] et de Bournaud [Bournaud 96] permettent de considérer des objets plus complexes. Toutefois, la hiérarchie résultante n'est pas destinée à être immédiatemment exploitée par l'utilisateur. Ce dernier doit auparavant s'attacher à élaguer lui-même la hiérarchie au moyen de mécanismes divers : Simon et Napoli proposent de calculer des points de vues sur la hiérarchie initiale, permettant d'observer plus finement tel ou tel phénomène ; Bournaud propose plusieurs algorithmes pour obtenir, soit automatiquement, soit manuellement, des hiérarchies moins complexe.
Dans notre cas, nous avons préféré décharger l'utilisateur du travail d'élagage de la hiérarchie. En effet, il ne s'agit pas ici de trouver une classification pertinente d'un certain nombre de phénomènes, mais de proposer une vision synthétique d'un ensemble d'informations. L'utilisateur ne doit pas passer du temps à comprendre la hiérarchie. Nos heuristiques sont le pendant des méthodes de simplification proposées par Bournaud et Simon. Elles sont moins souples et sans doute plus arbitraires, mais elles prennent en compte la spécificité des structures prédicatives. De plus, nous avons vu que le temps de calcul qui en découle est dans l'ensemble inférieur à celui du système COING. Notre méthode est donc moins générale, mais permet de gérer la complexité de la hiérarchie induite par l'utilisation de structures prédicatives, tant au niveau représentationnel qu'au niveau calculatoire.

Une comparaison plus poussée des différentes méthodes s'avère très difficile, car au delà des objectifs différents des travaux cités, la capacité d'une hiérarchie à structurer et organiser des connaissances ou informations est une notion très subjective. Nous sommes dans un domaine où l'évaluation d'un système est loin d'être évidente, car il n'y a pas de mesure objective, comme c'est le cas par exemple pour la prédiction de valeurs inconnues.

Les deux chapitres suivants (5 et 6) proposent une première évaluation de notre processus de généralisation. En effet, nous abordons la deuxième partie de notre mémoire, qui détaille l'application de notre méthode à l'analyse de l'information, et notamment une expérimentation sur un corpus de textes du domaine de l'agriculture. Les structures prédicatives sont utilisées pour améliorer une chaîne de traitement d'analyse de l'information. Par ailleurs, cette deuxième partie présente de manière plus approfondie l'aspect linguistique de notre travail.

5

Étude des travaux d'extraction d'informations à partir de textes pour l'analyse de l'information

Nous nous intéressons dans cette deuxième partie à l'analyse de l'information. Ce chapitre a pour but de présenter la notion d'analyse de l'information (section 5.1), puis de montrer comment on peut automatiser cette analyse, au moyens de méthodes et d'outils linguistiques et statistiques. Nous montrons ainsi l'intérêt d'une approche terminologique (section 5.2), puis nous intéressons aux relations entre termes et aux méthodes d'extraction existantes (5.3). Nous proposons enfin d'étendre une chaîne de traitement pour l'analyse de l'information, pour converger vers une méthode complète, intégrant notre processus de généralisation de structures prédicatives (section 5.4).

5.1 L'analyse de l'information pour caractériser un ensemble de documents

Les progrès technologiques dans le domaine de l'informatique et des télécommunications ont permis un accroissement sans précédent de la diffusion de l'information [Stephens 94]. En ce qui concerne plus particulièrement l'information à caractère scientifique et technique (articles, congrès, livres, manuels, ...), la masse croissante d'informations disponibles rend nécessaire l'utilisation de méthodes permettant de classer et de structurer le savoir. Ce problème est du ressort de l'informatique documentaire, qui propose des solutions pour la recherche d'informations dans des textes disponibles sous forme électronique[38].

L'informatique documentaire classique propose une méthode en deux étapes pour aider à trouver des informations pertinentes dans un ensemble de documents [Croft 92] :

- une première étape, l'indexation, consiste à assigner automatiquement des mots-clés ou descripteurs aux textes du corpus,

38. On utilisera le terme *corpus* pour désigner une collection de textes, supposés être disponibles sous forme électronique, et donc manipulables par des outils automatiques.

- une deuxième étape, la recherche d'informations, consiste à appliquer des stratégies automatiques de recherche de documents à partir d'une requête composée de descripteurs (combinaison booléenne, modèles probabilistes, ...).

L'emploi de quelques descripteurs pour décrire un texte n'autorise toutefois qu'une vue très partielle de ce dernier. Une approche alternative a été proposée, afin d'accéder de manière plus fine au contenu des textes : la recherche documentaire conceptuelle [Mauldin 91] [Van-Bakel 96]. Il s'agit d'effectuer une analyse automatique approfondie des textes et d'en déduire une représentation sémantique de leur contenu. Les requêtes subissent la même analyse, et sont mises en correspondance avec les représentations des textes. La recherche est ainsi effectuée de manière beaucoup plus complète. Cette approche nécessite une analyse aux niveaux syntaxique, sémantique et pragmatique des textes. Elle requiert des connaissances spécifiques nombreuses, d'ordre linguistique et conceptuel. Elle n'est donc applicable que sur des domaines très restreints, en dehors desquels elle s'avère infructueuse : d'une part les textes sont trop hétérogènes pour pouvoir être analysés correctement, d'autre part les sources lexicales et conceptuelles nécessaires sont inexistantes [Zweigenbaum 94].

La représentation du contenu des textes par des mots-clés est de fait surtout limitée par les analyseurs utilisés en informatique documentaire classique, qui intègrent peu de connaissances linguistiques. L'avènement de méthodes et d'outils d'analyse syntaxique partielle [Jacquemin 97, p.3] a permis d'améliorer sensiblement l'extraction de contenu avec des moyens restreints par rapport à ceux utilisés par l'approche conceptuelle : ainsi dans les domaines scientifiques et techniques, les extracteurs de termes permettent de prendre en compte de véritables unités d'information, généralement exprimées par des groupes nominaux, plutôt que de simple mots-clés. Comme le dit Jacquemin, il faut préférer « *une analyse fragmentaire et massives des textes pleins à une compréhension exhaustive en profondeur* ».

Ces nouvelles méthodes dites « faibles » par opposition aux approches complètes comme celle de Mauldin, ont permis le développement de travaux connexes à la recherche documentaire. Nous nous intéressons en particulier à ceux qui s'attachent à extraire et structurer le contenu informatif des textes. L'objectif est soit d'extraire des informations précises (Extraction d'Information), soit de synthétiser le contenu d'un ensemble de textes (Analyse de l'information).

L'Extraction d'Information

L'extraction d'informations précises à partir de textes est l'objectif des travaux en Extraction d'Information[39] [Cowie 96]. Les informations sont extraites pour remplir les champs d'une base de données définis très précisément. Les textes traités sont typiquement des articles de journaux ou de magazines. Le domaine d'application est souvent très restreint : par exemple, il peut s'agir de recenser les informations concernant les changements de postes de direction dans des entreprises. L'analyse des textes conduit à remplir des champs structurés par les informations appropriés. Par exemple, pour un article de journal commençant par la phrase (tiré de [Cowie 96]) :

> *A breakthrough into Eastern Europe was achieved by McDonald's, the American fast food restaurant chain ...*

39. De l'anglais *Information Extraction*.

les champs suivant sont remplis par le système d'analyse :

ENTITY-1375-12>:=
NAME : MCDonald's
NATIONALITY : U.S.
TYPE : Company

L'extraction d'information fait l'objet de nombreuses recherches, une conférence annuelle permettant de tester les systèmes existant (MUC, Message Understanding Conference). Les systèmes sont généralement conçus comme une suite de composants linguistiques spécialisés, allant du filtrage statistique à l'analyse de la référence en passant par une analyse syntaxique partielle et une analyse sémantique. Ils sont généralement très dépendants du domaine visé. Un des résultats majeurs de ces travaux est que les systèmes privilégiant une approche « faible » sont plus efficaces que les systèmes développant des grammaires ou des modèles de discours sophistiqués [Salton 94]. L'extraction d'information n'est toutefois applicable que sur des domaines très restreints et des textes homogènes, pour les mêmes raisons que pour la recherche documentaire conceptuelle.

L'Analyse de l'Information

L'extraction et la structuration du contenu informatif d'un corpus de textes, afin d'en obtenir une caractérisation synthétique, est l'objectif de l'Analyse de l'Information [Muller 97]. C'est une notion récente, qui ne constitue pas encore un champ de recherches bien structuré comme l'Extraction d'Information. Nous retiendrons la définition qui en est donnée par Toussaint et al. [Toussaint 98] :

> « *L'analyse de l'information peut être définie comme un ensemble d'outils et de méthodes permettant à un opérateur humain de collecter l'information contenue dans un corpus sans le lire de façon séquentielle.* »

L'analyse est en particulier destinée à des fonds documentaires à caractère scientifique et technique. Elle doit permettre d'identifier l'information utile, comportant un intérêt pour l'utilisateur.

De façon plus concrète, nous pouvons présenter deux situations où intervient l'analyse de l'information. En premier lieu, celle-ci peut être vue comme l'étape intervenant après une recherche d'informations dans un fonds documentaire [Toussaint 96]. La recherche d'informations, nous l'avons vu, consiste à interroger une base documentaire. Le système de recherche fournit une réponse sous forme de liste de documents et de leur résumé. Lorsqu'il y a beaucoup de réponses, examiner l'ensemble des documents peut prendre beaucoup de temps. Le processus d'analyse de l'information permet de simplifier ce traitement en caractérisant le contenu de l'ensemble des documents et en fournissant une représentation synthétique de l'information correspondante.

En second lieu, l'analyse de l'information peut être utilisée dans le cadre de la veille technologique. Par exemple, l'INIST[40] est parfois amené à constituer des rapports de tendances sur l'évolution d'un domaine de connaissances : il s'agit de mettre en évidence les thèmes traités sur ce domaine, ainsi que les acteurs concernés, à partir d'un fonds documentaire. Récemment, l'INIST a ainsi fourni un rapport sur les plantes transgéniques. L'analyse automatique est dans ce cas beaucoup moins précise que celle réalisée pour

40. Institut National de l'Information Scientifique et Technique, Vandoeuvre (54).

l'Extraction d'Information, car elle ne remplit pas des champs de données précis. Elle permet de couvrir des domaines plus vastes, moins spécialisés. Elle sert de support au travail final de synthèse réalisé par des documentalistes experts d'un domaine.

Nous montrons dans les sections suivantes (5.2 et 5.3) les différentes méthodes et outils qui peuvent être utilisés en analyse de l'information, en insistant tout d'abord sur la nécessité d'une approche terminologique. Puis nous discutons d'une des premières approches utilisée pour l'analyse de l'information, qui combine des méthodes linguistiques et statistiques [Toussaint 97], et proposons une extension prenant en compte les structures prédicatives (section 5.4).

5.2 L'approche terminologique pour extraire l'information de textes scientifiques

Les notions de mots-clé ou de descripteur utilisées en recherche d'information depuis très longtemps sont souvent réductrices. Elles représentent les unités lexicales qui sont utilisées comme accès à l'information, sans considérer leur dimension linguistique. Il s'agit en fait, la plupart du temps, de termes. La terminologie, dont l'objet est l'étude des termes, est un passage obligé pour une meilleure prise en compte des entités de base qui véhiculent l'information. C'est ce que nous montrons dans cette section (section 5.2.1). Nous étudions ensuite quelques unes des méthodes permettant l'identification automatique des termes dans les textes (section 5.2.2).

5.2.1 La terminologie et l'information

La terminologie est une discipline dérivée de la linguistique, datant du milieu de ce siècle [Felber 87]. Par *terminologie*, on peut en fait désigner trois notions distinctes : une science fortement linguistique ; un ensemble de méthodes pour l'étude et le classement des termes d'un domaine de connaissances ; l'ensemble des termes d'un domaine de connaissances [Sager 90]. Les deux notions principales associées à la terminologie sont le *terme* et le *domaine*.

Le terme peut être défini comme « *une unité signifiante constituée d'un mot* [terme simple] *ou de plusieurs* [terme complexe] *qui désigne une notion univoque à l'intérieur d'un domaine* » [Viallet 94]. Par exemple, *menu* est un terme simple dans le domaine de l'informatique. De même, *langage orienté objet* est un terme complexe dans le domaine de l'informatique. Les termes sont la plupart du temps des groupes nominaux, mais peuvent également être des verbes ou des adjectifs.
Nous voyons qu'un terme se distingue d'un mot de la langue courante par des caractéristiques théoriques bien précises, notamment [Sager 90] :

- il désigne une et une seule notion,
- il possède une efficacité fonctionnelle marquée, sans connotation émotive,
- il est peu dépendant du contexte, c'est-à-dire que sa signification est stable, quel que soit le contexte d'utilisation du terme,
- il est lié à un domaine (technique, science), en dehors duquel il n'a pas de signification,
- il évolue peu.

Au vu de ses propriétés, la tentation est grande de considérer un terme comme une simple étiquette apposée sur un concept. Cette vue est toutefois démentie par la pratique, et critiquée par de nombreux auteurs ([Rastier 95], [Skuce 91] [Condamines 97], [Jacquemin 97]). Les termes sont des expressions linguistiques à part entière, et leur complexité doit être prise en compte. Ainsi, de nombreux phénomènes linguistiques se retrouvent parmi les termes : les homonymies (A et B ont la même graphie mais des sens totalement différents), la polysémie (A et B ont la même graphie, mais possèdent des sens voisins), les variations morphologiques et syntaxiques (A et B sont des expressions linguistiques différentes mais de même sens). Les variations sont particulièrement sous-estimées. Jacquemin a pourtant montré qu'elles peuvent représenter jusqu'à 25% des occurrences de termes pour la langue française [Jacquemin 97].

A mi-chemin entre la fluidité d'un mot de la langue courante, dont le sens est bien souvent multiple[41], et l'immuabilité et la précision d'une étiquette, le terme est recherché parce qu'il est le support privilégié de l'information contenu dans un texte. Cette propriété découle naturellement de sa définition, clairement exprimée par Jacquemin [Jacquemin 97] :

> « *Les langues de spécialités* [relatives à un domaine de spécialité] *sont le support de la communication scientifique et technique. L'information y est contenue en priorité dans les termes qui sont la forme linguistique des concepts du domaine.* »

Les textes scientifiques et techniques font un usage élevé de termes, car ils recherchent *a priori* la précision et l'objectivité des propos, et non l'émotion de leurs lecteurs. C'est pourquoi une approche terminologique est adaptée à des textes scientifiques, mais non à des textes de la langue courante.

Les travaux de Harris sur la notion de sous-langage (analogue aux langues de spécialités[42]) permettent d'approfondir la notion de contenu informationnel d'un texte [Harris 89]. Présentés de façon schématique, ses travaux se sont attachés à réduire l'information contenue dans un texte à un ensemble de formules exprimés à partir d'un ensemble d'opérateurs et d'arguments : les phrases sont décomposées en syntagmes noyaux, unités d'information minimales du corpus correspondant aux termes. Les formules expriment des contraintes sur la combinatoire des opérateurs et arguments. Les opérateurs et arguments sont regroupés en classes distributionnelles, c'est-à-dire selon leurs positions possibles dans la structure d'une phrase.
Par exemple, les huit phrases du tableau 5.1, tiré des travaux de Harris et al. sur la médecine [Harris 89], possèdent le même contenu informationnel, exprimé par la formule :

$$A\ V_p\ C_z$$

V_p est la classe des opérateurs qui expriment la notion de *production*. La classe V_p est une sous-classe des opérateurs V exprimant la notion de *réponse*. V_p regroupe des termes divers tel que *formation*, *production*, *synthèse*, *résultat*. A est un argument qui exprime la classe des *anticorps*. C_z est un argument qui exprime la classe des *plasmocytes*, qui est une sous-classe de C, les cellules.

Les classes sont ainsi constitués par un ensemble de termes ayant un contenu informationnel identique ou voisin. Les formules de Harris permettent de factoriser les multiples

41. Essayer de compter par exemple les sens d'un mot comme *marché* ou *bulletin* ...
42. Pour une vue générale des langues de spécialités, voir [Lerat 95].

1	Les plasmocytes sont producteurs d'anticorps.
2	Les plasmocytes produisent des anticorps.
3	Des anticorps sont produits par les plasmocytes.
4	Des anticorps sont produits dans les plasmocytes.
5	La production plasmocytaire d'anticorps a été établie.
6	La production d'anticorps par les plasmocytes a été établie.
7	La production d'anticorps qu'on observe dans les plasmocytes, a été établie.
8	L'origine plasmocytaire de la production d'anticorps a été établie.

TAB. 5.1 – *Huit phrases de contenu informationnel identique $AVPC_z$ d'après [Harris 89]*

expressions linguistiques pouvant correspondre à une information. Le travail réalisé par Harris et al. sur des textes de médecine est une formidable tentative de formalisation du contenu. Il reste toutefois hors de portée d'un système automatique, car il requiert des transformations complexes des phrases initiales pour aboutir à un ensemble de formules. Notons que la représentation que nous utilisons pour la généralisation est très voisine de celle-ci, puisque nos prédicats correspondent aux opérateurs de Harris, et que les classes sont également hiérarchisées. Toutefois, une différence importante est que les classes de Harris sont basées sur la distribution des termes dans les textes, et non sur des propriétés sémantiques comme une hiérarchie conceptuelle.

Cet exemple nous a permis d'illustrer la variété linguistique pour exprimer une information. Pour saisir le contenu informationnel d'un texte, il faut donc être capable rapprocher des expressions différentes mais conceptuellement liés. La prise en compte de la variation terminologique permet de résoudre une partie du problème, comme nous le verrons plus loin avec le système FASTER (section 5.2.2).

Le domaine et les relations entre termes

Le domaine est une notion qui, paradoxalement, est mal définie. C'est pourtant un des fondements de la théorie terminologique [Otman 94] : le domaine représente un espace conceptuel structuré, délimitant une discipline ou une technique, et dont la visibilité passe par les termes. Au sein d'un domaine, les termes ont une forte cohésion, structurés par un ensemble de relations les situant les uns par rapport aux autres. Cette vision des choses peut être directement comparée aux travaux en intelligence artificielle sur les bases de connaissances. De fait, on retrouve en intelligence artificielle, dans les bases de connaissances ou les ontologies[43] relations qui sont utilisées depuis longtemps en terminologie, la plus connue étant la relation de spécialisation, appelée *hypéronymie* en terminologie [Otman 93]. Un panorama des relations utilisées de part et d'autres (par exemple, [Gouadec 94] pour la terminologie, [Nutter 89] pour l'intelligence artificielle) montre une approche similaire pour structurer la connaissance. Il ne faut toutefois pas tomber dans un excès de simplification. Les significations des relations utilisées peuvent être très différentes : les relations entre concepts ne sont pas les mêmes que les relations entre unités lexicales. Nous pouvons reprendre l'exemple de la relation de spécialisation, qui est sans doute la relation ayant fait l'objet des plus nombreux travaux, aussi bien en représentation des

43. Une ontologie peut être définie comme la spécification du vocabulaire conceptuel d'un domaine de discours, ou comme un modèle des catégories de connaissances d'un domaine [Rector 96] [Reynaud 94]. Des exemples bien connus d'ontologies sont CYC [Guha 90] et UMLS [Lindberg 93].

connaissances qu'en terminologie ou en sémantique lexicale , : elle correspond dans le cas de la subsomption dans les logiques de descriptions à une relation formalisée, pouvant être définie de manière structurelle, intensionnelle ou extensionnelle [Woods 91] ; dans le cas de la terminologie, elle correspond à une inclusion de propriétés [Lerat 90] ; dans le cas de la sémantique lexicale, elle est définie par rapport à des tests lexicaux (X est un hypéronyme de Y si la phrase *c'est un X* implique la phrase *c'est un Y*) [Cruse 86].

En dehors de sa visée théorique, la terminologie est aussi une pratique qui s'attache à recenser les termes et leur comportement dans tous les domaines de spécialités. Les termes sont reconnus comme tels sur la base de trois critères [Orlac 94] : la fréquence, la syntaxe de la phrase, la place du candidat dans un paradigme (c'est-à-dire dans le système notionnel représentant le domaine). Toutefois, ces critères sont approximatifs, et le choix repose surtout sur l'intuition des terminologues, chargés de leur reconnaissance [Bourigault 95]. Le recensement des termes d'un domaine conduit à la construction d'une base de données terminologique, collectant des informations de nature conceptuelle (définition, relations avec les autres termes), linguistique (graphie, variantes) et pragmatique (contextes d'emploi, usage) sur les termes.

Les chercheurs en intelligence artificielle ont été nombreux à voir l'intérêt de la terminologie pour les applications sur l'acquisition de connaissances, le langage naturel, et de manière indirecte, pour la recherche d'informations [Meyer 91], [Meyer 94], [Aussenac-Gilles 95], [Condamines 92], [Otman 94]. Des systèmes d'extraction de termes ont vu le jour, afin d'automatiser les processus d'extraction et de reconnaissance. Ces systèmes prennent en compte les phénomènes linguistiques, et permettent d'obtenir de bien meilleurs résultats que les premiers systèmes d'indexation automatique négligeant les connaissances linguistiques. Nous présentons dans la section suivante différents systèmes d'extraction de termes, qui montrent les possibilités actuelles en la matière.

5.2.2 L'acquisition et la reconnaissance de termes

De nombreux travaux ont vu le jour pour automatiser l'extraction de termes dans les textes ([Bourigault 94a], [Dumas 96], [Oueslati 96], [Herviou 95]). On distingue généralement l'*acquisition*, qui consiste à générer une liste de candidats termes à partir d'un corpus de textes, et la *reconnaissance*, qui consiste à repérer dans les textes un ensemble de termes fournis a priori.

L'acquisition terminologique est destinée à la création ou à l'augmentation d'une terminologie. L'objectif est de constituer la liste des termes liés à un domaine, et éventuellement de les organiser. Les systèmes correspondant fournissent de fait des *candidats termes*, qui doivent par la suite faire l'objet d'une validation humaine qui conduit généralement au rejet d'un bon nombre de candidats.

La reconnaissance terminologique est orientée vers l'indexation, et se fixe comme objectif de repérer des occurrences de termes et de leur variantes. Elle permet de se focaliser sur une liste précise de termes et s'avère plus pertinente pour extraire le contenu des textes : le bruit généré est beaucoup moins important que dans le cas de l'acquisition terminologique.

Trois grandes approches sont utilisées, que ce soit pour l'acquisition ou la reconnaissance terminologique : celle utilisant des méthodes linguistiques, celle utilisant des méthodes statistiques, et celle couplant les deux types de méthodes [Jacquemin 97]. Nous présentons ici trois systèmes d'extraction terminologique : LEXTER utilise une approche linguistique

pour l'acquisition ; ACABIT utilise une approche mixte pour l'acquisition ; FASTER utilise une approche linguistique pour la reconnaissance. Les approches purement statistiques sont peu adaptées aux français et ne seront pas abordées dans ce mémoire (pour un panorama complet des travaux et systèmes, on pourra se reporter à [Jacquemin 97]).

Avant de détailler ces systèmes particuliers, nous présentons les étapes préparatoires nécessaires au traitement des corpus. Presque tout système traitant des textes sous forme électronique doit en effet procéder à un pré-traitement de ces textes, qui consiste généralement en :

- une segmentation du texte en phrase. Même si à première vue cette étape semble facilement automatisable, certains phénomènes (abréviations, noms propres, ...) rendent difficile un découpage automatique parfait,
- un étiquetage des mots par leur catégorie grammaticale : déterminants, noms communs, prépositions, adjectifs, ...,
- une lemmatisation des mots fléchis, qui consiste à associer à chaque forme fléchie, c'est-à-dire à chaque occurrence d'un mot dans le texte, un lemme et un ensemble de traits flexionnels (temps, genre, nombre).

Ce pré-traitement est essentiel car il permet de normaliser les occurrences de mots dans les textes, et d'ôter les ambiguïtés pour certaines formes lexicales. L'étiquetage est relativement bien maîtrisé, notamment par des systèmes probabilistes à base de règles tel que l'étiqueteur de Brill [Brill 93]. Celui-ci, après une phase d'entraînement sur une partie étiquetée manuellement du corpus de textes à traiter, permet d'obtenir des taux de réussites supérieurs à 95%. Pour une langue comme le français, la lemmatisation est difficile, car la morphologie flexionnelle est complexe (modèles de conjugaison, familles de flexions). Les systèmes automatiques de lemmatisation peuvent être basés sur la consultation d'un dictionnaire, ou sur un ensemble de règles, voire sur une combinaisons des deux. L'avantage des systèmes basés sur les règles est la prise en compte de mots inconnus du dictionnaire et la détection de certaines erreurs d'étiquetage. Le pré-traitement permet ainsi de bien identifier et séparer les unités lexicales qui vont être manipulées par la suite [Fuchs 93].

Le système LEXTER : méthode linguistique pour l'acquisition

Le système LEXTER [Bourigault 94a] [Bourigault 94b] utilise une méthode linguistique et est destiné à l'acquisition. Il procède par découpage au sein des phrases d'un texte : un ensemble de patrons syntaxiques permettent de repérer les portions qui ne peuvent être des constituants de termes et signalent ainsi les frontières des termes. Les ambiguïtés dues à différentes possibilités d'application des règles de découpage sont résolues par des procédures dites endogènes, permettant de chercher par ailleurs dans le texte une situation équivalente et non ambiguë. Les groupes nominaux obtenus sont ensuite décomposés en deux constituants, de manière récursive. Ceci permet d'obtenir des liens entre candidats termes, et de constituer un réseau terminologique accessible au moyen de fonctions hypertextes. Par exemple, étant donné la phrase suivante :

> *L'alimentation en eau est assurée par une pompe d'extraction, qui est reliée à l'alimentation électrique de la pompe de refoulement,*

trois groupes nominaux sont extraits :

> *alimentation en eau,*

pompe d'extraction,
alimentation électrique de la pompe de refoulement.

Le premier candidat terme peut être décomposé en une tête, *alimentation* et une expansion, *eau*. Cette décomposition permettra par exemple de le relier par exemple au candidat *alimentation en huile* par l'intermédiaire de sa tête. Le troisième candidat terme peut être décomposé en *alimentation électrique* et *pompe de refoulement*, eux-même décomposés à leur tour en *alimentation* et *électrique* et *pompe* et *refoulement*. Il sera également lié au premier candidat par l'intermédiaire de sa tête *alimentation*.

LEXTER extrait un grand nombre de candidats termes à partir d'un corpus de textes. Il nécessite une vérification manuelle qui peut s'avérer fastidieuse. La présence d'un module hypertexte d'exploration permet toutefois une première structuration bien utile, plus pertinente qu'une simple liste. De plus, des modules de structuration supplémentaires ont été conçus pour regrouper des termes selon des régularités syntaxiques [Assadi 96]. Ces derniers ont été utilisés en acquisition des connaissances [Assadi 97].

Le système ACABIT : approche mixte pour l'acquisition

Le système ACABIT [Daille 94], destiné à l'acquisition, utilise une approche mixte, linguistique et statistique. Il effectue tout d'abord une analyse par automates, permettant de repérer les séquences de mots ayant une structure syntaxique caractéristique des termes. Les séquences détectées sont des noms composés binaires, constitués de deux mots qui ne sont pas fonctionnels[44]. Ces filtres linguistiques sont capables de prendre en compte certaines variations. Par exemple, la séquence :

système racinaire de surface et de profondeur

peut être reconnue comme l'occurrence de deux termes, *système de surface* et *système de profondeur*, par la règle suivante :

nprep_ncoordN → nprepn COORD PREP NCOMM

nprepn_ncoordN représente la séquence entière, décomposable comme suit : *nprepn* représente deux noms communs séparés par une préposition (*système de surface*), COORD une coordination (*et*), PREP une préposition (*de*) et NCOMM un nom commun (*profondeur*). ACABIT effectue ensuite un filtrage statistique sur les candidats termes binaires obtenus. Pour chaque phrase, deux candidats termes sont retenus sur la base d'un coefficient de vraisemblance qui classe les séquences selon un ordre de pertinence.

ACABIT, comme LEXTER, permet de recueillir un nombre important de candidats termes à partir d'un corpus de textes, et nécessite une vérification manuelle importante. L'importance de l'information apportée par tous les candidats est en effet très variable, car non contrôlée.

Le système FASTER : reconnaissance par méthode linguistique

Le système FASTER [Jacquemin 95] [Jacquemin 97] est prioritairement destiné à la reconnaissance de termes, bien qu'il puisse être également utilisé dans le cadre de l'acquisition. FASTER, contrairement aux autres approches, utilise une liste de termes prédéterminée. Dans ce cas, on parle d'indexation contrôlée, car les termes à retenir ont été

44. Les mots dits fonctionnels sont les prépositions, articles, conjonctions, ...

donnés par avance. Le système ne se contente toutefois pas de rechercher les occurrences des termes de la liste : il met en oeuvre un mécanisme puissant de recherche des variations des termes, qui permet de prendre réellement en compte la complexité linguistique des termes.

FASTER utilise une métagrammaire de la langue à traiter, qui décrit des patrons de variation terminologique susceptibles d'être rencontrés (environ une centaine pour le français). Cette grammaire est formalisée au moyen de métarègles qui expriment les variations, qui peuvent être d'ordre morphologique ou syntaxique.
Les variations morphologiques sont :

- les flexions : singulier/pluriel, infinitif, participe passé,.... Par exemple, en agriculture, le pluriel *amines* est une variation par flexion du terme *amine*,
- les dérivations : passage d'un nom à un adjectif, d'un verbe à un nom (nominalisation). Par exemple, en agriculture, le groupe nominal *système racinaire* est une variante dérivationnelle du terme *système de racine*.

Les variations syntaxiques sont les suivantes :

- l'insertion ou modification est l'introduction d'un mot non fonctionnel à l'intérieur d'un groupe nominal. Par exemple, en informatique, le groupe nominal *approche orientée objet* est une variation par insertion du terme *approche objet*,
- la coordination est une forme coordonnée de mots (adjectifs, noms) l'intérieur d'un groupe nominal. Par exemple, en médecine, le groupe nominal *artères bronchiales et intercostales* est une variation par coordination du terme *artères bronchiales*,
- la permutation implique un élément pivot autour duquel les mots ou groupes de mots peuvent permuter. Par exemple, en médecine, pour la langue anglaise, le groupe nominal *dissemination in blood* est une variante de permutation du terme *blood dissemination*.

Le système peut également prendre en compte des variations morpho-syntaxiques, qui combinent les deux types de variations. Les meta-règles sont exprimées à l'aide d'un formalisme de structure de traits. Voici par exemple une règle pour traiter un cas simple de permutation en anglais :

Perm(X1 → X2 X3) = X1 → X3 P4 X2
<X2 cat>≠ P
<P4 lemme>= 'of'
<X2 cat>≠ A
<X3 cat>≠ P

La partie gauche de la première ligne décrit la structure d'un terme X1 qui peut se réécrire en X2 X3. Le symbole X correspond à un mot de catégorie quelconque. Les symboles P et A sont respectivement utilisés pour désigner des prépositions et des adjectifs. La partie droite de la première ligne montre que le terme X1 peut accepter une autre structure où X2 et X3 sont permutés et séparés par la préposition *of*. Les lignes suivantes représentent des traits associés aux différents mots : le trait *cat* représente la catégorie syntaxique, le trait *lemme* la racine du mot. Ces traits représentent des contraintes sur les mots : X2, par exemple, ne peut être ni une préposition, ni un adjectif. Cette règle est capable par exemple de reconnaître la variante *planting of seed* à partir du terme *seed planting*, avec X2 = *seed* et X3 = *planting*.

FASTER est un outil qui permet de prendre en compte la nature complexe des termes, en traitant les variantes susceptibles d'être rencontrées dans les textes. Il contribue à une amélioration notable de la recherche de descripteurs ou mots-clés, utilisée pour l'informatique documentaire [Daille 96]. Il permet d'effectuer une indexation contrôlée, qui semble être la meilleure solution pour extraire des textes un contenu informatif de qualité. Il peut de plus s'appliquer sur différentes langues, selon l'ensemble de meta-règles utilisées. L'inconvénient est qu'il faut disposer a priori d'une liste présentant une bonne couverture des termes présents dans le corpus de textes à traiter. Une approche mixte est alors possible, utilisant un système d'acquisition en premier lieu pour constituer une liste de termes suffisamment complète, et le système FASTER par la suite. Il est toutefois nécessaire de contrôler manuellement la liste de termes issue du système d'acquisition.

Conclusion sur les systèmes d'extraction terminologique

Les systèmes d'extraction terminologique sont indispensables pour repérer les unités d'informations dans un corpus de textes à caractère scientifique ou technique. Dans une perspective d'identification, une système comme FASTER permettant de repérer les variantes de termes et d'opérer une certaine normalisation semble la meilleure solution. Bien sûr, FASTER est loin de pouvoir identifier comme une unique information les huit phrases données en exemple dans la section 5.2.1, car il opére au niveau des groupes nominaux et non des phrases entières. La prise en compte de structures plus importantes, jusqu'à la taille d'une phrase, requiert la prise en compte des relations entre les différentes unités d'informations que sont les termes. Nous étudions dans la section suivante les différents travaux sur l'identification et la caractérisation de ces relations.

5.3 Les relations entre termes, pour structurer les unités d'informations

Nous nous intéressons principalement aux relations susceptibles d'apparaître dans les textes, et qui peuvent faire l'objet d'une extraction. Nous ne prétendons pas ici faire le tour complet des recherches dans ce vaste domaine qu'est l'acquisition automatique d'informations lexicales à partir de corpus[45] : nous nous focalisons sur certains travaux permettant d'identifier des relations entre unités lexicales[46], afin de mettre à jour de manière plus complète le contenu informationnel des textes. Il s'agit donc de dépasser le simple niveau des termes pour atteindre des structures syntaxiques plus vastes, mettant en jeu plusieurs termes au sein des phrases.

Dans cette section, nous faisons tout d'abord une distinction entre deux grands types de relations, syntagmatiques et paradigmatiques, et montrons que nous sommes principalement intéressés par les relations syntagmatiques (section 5.3.1). Puis nous nous intéressons à deux types d'approches permettant d'identifier des relations :

- les approches linguistiques statistiques, dont la mise en oeuvre est simple et les connaissances linguistiques nécessaires limitées ; elles permettent d'obtenir des re-

45. [Boguraev 96] et [Pichon 97] présentent une bonne vue d'ensemble de ce domaine.

46. Par défaut, les unités lexicales considérées sont des mots. Nous précisons le cas échéant si les travaux s'appliquent plus particulièrement à des termes.

lations entre mots ou des classes de mots, sans toutefois pouvoir préciser la nature ou le sens des relations mises en jeu [Grefenstette 94],
- les approches symboliques, dont la mise en oeuvre est plus complexe et nécessite de vastes connaissances linguistiques ; les relations entre les mots sont représentées au sein de structures prédicats-arguments (structures prédicatives) qui rendent compte de manière sophistiquée du contenu informationnel [Saint-Dizier 95].

Les sections 5.3.2 et 5.3.3 sont respectivement consacrées à une présentation critique de travaux correspondants à ces deux types d'approches.

5.3.1 Deux grands types de relations

Les relations entre unités d'informations peuvent être de nature très différentes. La sémantique lexicale retient deux grands type de relations : syntagmatiques et paradigmatiques [Lyons 77, p. 240-241].

Les relations syntagmatiques résultent de l'association de différents syntagmes dans une même phrase, c'est-à-dire de différentes unités qui peuvent faire partie du même voisinage syntaxique. Par exemple, l'expression *dosage des amines* fait apparaître une relation syntagmatique entre les deux termes *dosage* et *amine*.

Les relations paradigmatiques lient des unités qui sont substituables les unes aux autres dans un même contexte syntaxique, c'est-à-dire qui ont un rôle similaire dans une position syntaxique identique. Par exemple, les expressions *dosage des amines* et *dosage des polyamines* font apparaître une relation paradigmatique entre les deux termes *amine* et *polyamine*.
Les relations paradigmatiques sont majoritairement celles qui regroupent les unités selon leur similarité sémantique : toutes les unités X pouvant apparaître dans l'expression *dosage de X* font partie d'une même classe sémantique. Ce raisonnement est toutefois approximatif, car il n'existe pas une parfaite symétrie entre syntaxe et sémantique [Habert 96b]. On peut trouver des unités X qui ne sont pas des substances chimiques, classe sémantique que l'on voudrait associer à ces unités. Par exemple, *dosage de l'activité des amines* met en oeuvre une expression plus complexe. Bouaud et al. ont toutefois montré, en comparant des classes syntaxiques paradigmatiques et des classes sémantiques formées sur un même domaine, que les unités d'informations sont regroupées selon leur proximité sémantique, même si les critères divergent [Bouaud 97]. Les relations paradigmatiques sont à rapprocher des relations structurelles organisant les bases de connaissances ou les ontologies : les relations de spécialisation et partie-tout, qui permettent de définir entre elles les différentes unités (concepts ou items lexicaux). Par exemple, la relation entre *polyamine* et *amine* est une relation de spécialisation, qui permet de définir *polyamine* par rapport à l'entité plus générale *amine*.

Au contraire, les relations syntagmatiques expriment les associations « accidentelles » entre les unités d'informations, qui constituent la substance d'un discours sur un domaine. Ce sont ces relations qui nous intéressent le plus, car elles véhiculent le contenu informatif du texte. Ainsi, l'association entre *dosage* et *amine* nous donne une information sur un événement et l'objet associé dans un texte donné, et non une connaissance définitoire.

Nous verrons par la suite que les deux types de relations peuvent être identifiés dans les textes : ce sont toutefois les relations syntagmatiques qui prédominent et qui sont les plus intéressantes pour l'analyse de l'information.

5.3.2 Les méthodes linguistiques statistiques : classes de mots

Pour rechercher des relations entre les mots, la méthode la plus simple est de considérer la proximité syntaxique de ceux-ci, en se basant sur le constat suivant : deux mots qui apparaissent souvent dans une même phrase sont probablement liés sémantiquement [Habert 97]. Cette constatation ouvre la voie à toute une série de méthodes, dont les deux principaux paramètres sont les suivants : le contexte du mot et la fonction statistique utilisée pour décider de la proximité des mots considérés.

Le contexte du mot correspond à son entourage syntaxique. Pour calculer les mots qui sont en lien avec un mot m, plusieurs contextes différents sont possibles [Grefenstette 93] :

- la phrase entière dans laquelle apparaît m (on dit que les mots sont co-occurrents),
- une fenêtre de n mots de part et d'autre de m (n est alors un paramètre supplémentaire),
- les unités lexicales voisines de m appartenant à une catégorie syntaxique donnée (nom, adjectif, verbe) ou entretenant une relation syntaxique particulière avec m (sujet d'un verbe, objet d'un verbe).

Parmi ces contextes, les mots fonctionnels sont généralement négligés. Les contextes définis au moyen de relations syntaxiques conduisent à des méthodes plus complexes à mettre en oeuvre, car elles nécessitent l'utilisation d'analyseurs syntaxique de surface.

Plusieurs mesures statistiques peuvent être utilisées pour calculer le poids relatif de la relation entre deux mots. La plus connue est sans doute celle de l'information mutuelle [Church 90], qui s'exprime par la formule :

$$MI(x, Y) = log(P(x, y)/P(x)P(y))$$

où $P(x, y)$ représente la probabilité que x et y co-occurrent, et $P(x)$ la probabilité que x occurre. L'information mutuelle est maximale pour deux mots qui apparaissent toujours ensemble, et qui sont susceptibles d'être fortement liés.

Nous allons présenter plus particulièrement quatre approches différentes pour identifier des relations entre unités lexicales. Ces quatre approches ont en commun de s'intéresser aux langues de spécialités et donc à des termes plutôt qu'à de simples mots comme c'est le cas pour Grefenstette [Grefenstette 93] ou Church et Hanks [Church 90].

Les travaux de Assadi et Bourigault : classification de noms et d'adjectifs

Nous avons déjà mentionné ces travaux en présentant le système d'extraction terminologique LEXTER (cf. section 5.2.2). A partir du réseau grammatical construit par LEXTER, les auteurs proposent une méthode pour constituer des classes d'adjectifs et de noms apparaissant dans les même contextes [Assadi 96], [Assadi 97]. L'objectif est de construire un système de consultation de documents techniques qui permette à un utilisateur de naviguer au sein de la documentation à l'aide d'un index ou des concepts du domaine.

Les candidats termes expriment des contextes terminologiques pour un groupe nominal donné. Par exemple, le groupe nominal *réseau* possède le contexte {*national, régional, distribution*} grâce aux termes {*réseau national, réseau régional, réseau de distribution*}. L'idée est de créer des classes de groupes nominaux ayant des contextes terminologiques similaires. Pour cela, chaque groupe nominal se voit associer un vecteur d'attributs qui

Classe	Contexte
antenne, barre, cable, cable souterrain, liaison, ligne, niveau, ouvrage, ...	HT, THT, tension, souterrain, MT, simple, servie, haute tension, prévu
composante, courant, courant de court-circuit, court-circuit, hydraulicité, intensité, longueur, perturbation, puissance, ...	maximal, supérieur, inverse, nominal, maximum, admissible, harmonique, direct, minimal, réactif, secondaire, moyen, ...

TAB. 5.2 – *Deux classes et leur contextes issues du réseau grammatical de LEXTER d'après [Assadi 97]*

décrit son contexte. Une classification hiérarchique ascendante est ensuite effectuée sur ces vecteurs : les deux groupes nominaux partageant le plus proche contexte sont d'abord regroupés, puis le processus est réitéré jusqu'à regrouper l'ensemble des groupes nominaux. L'arbre de classification ainsi obtenu est ensuite coupé à un certain niveau pour former des classes. Ce processus s'applique aussi bien pour des contextes constitués par des adjectifs que pour des contextes constitués par des autres groupes nominaux. La table 5.2 montre par exemple deux classes et leur contexte dans le domaine de la planification de réseaux électriques régionaux, extraits de [Assadi 97].

Cet outil de classification est complété par des outils de typage conceptuel avec l'aide du réseau grammatical : détermination de liens *sorte-de*, de liens *objet-attribut* et *objet-action*. Ces travaux s'inscrivent dans une méthode globale d'acquisition de connaissances à partir de textes qui attachent une attention particulière aux problèmes linguistiques et à la difficulté du passage des termes aux concepts [Bourigault 94b] [Assadi 96]. Ils proposent donc des outils d'exploration nécessitant un travail important de modélisation de la part de l'utilisateur du système, qui s'avère peu adapté à la caractérisation directe du contenu des textes.

Les travaux de Habert et al. : un réseau lexical pour présenter une image réorganisée du texte

Les travaux de Habert et al. [Habert 96a] [Habert 96b] entretiennent un étroit rapport avec les travaux de Assadi et Bourigault, puisqu'ils utilisent le même point de départ, à savoir le réseau grammatical fourni par le système LEXTER. L'objectif est, dans une perspective Harissienne (cf. section 5.2.1), de mettre à jour les classes d'opérateurs et d'arguments du domaine de spécialité considéré. Ceci afin de proposer un réseau lexical pouvant constituer une image réorganisée du texte initial, utilisé pour l'acquisition de connaissances à partir de textes.

La méthode proposée consiste dans un premier temps à simplifier les groupes nominaux extraits par LEXTER, de la même façon que Harris simplifie les phrases pour obtenir des phrases élémentaires mettant à jour les opérateurs et arguments d'un sous-langage. Cela permet d'obtenir des arbres élémentaires, qui mettent en évidence des collocations entre différentes unités lexicales. Ces collocations sont utilisées pour constituer des classes de contextes syntaxiques, à la manière des travaux de Assadi et Bourigault. Toutefois, la différence se situe dans l'exploitation qui en est faite : les classes sont directement utilisées pour construire une visualisation graphique sous forme de réseau lexical, montrant les

connexions entre unités lexicales. La figure 5.1 reproduit un tel réseau[47].

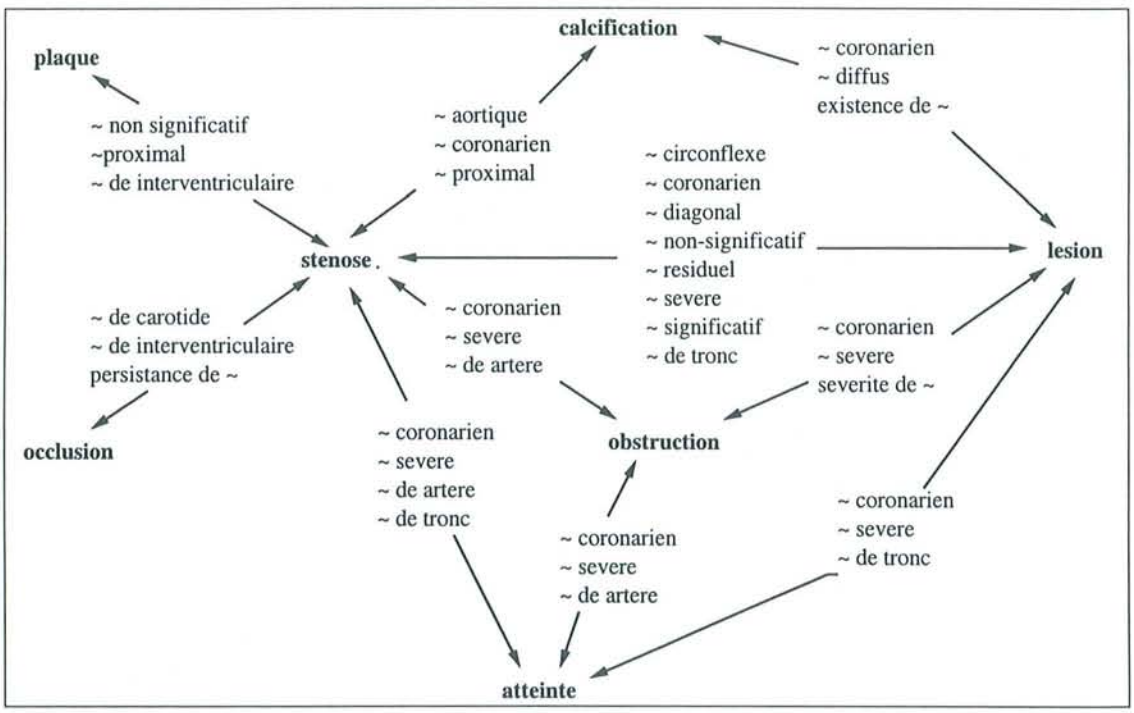

FIG. 5.1 – *Un réseau lexical représentant les affections corporelles localisées d'après [Habert 96b]*

Habert et al. se situent eux aussi dans un perspective d'acquisition de connaissances. Ils montrent bien que les travaux d'extraction à partir de textes, basés sur la syntaxe, achoppent sur le passage à une représentation conceptuelle : les classes mises à jour ne reflètent pas forcément la structure conceptuelle du domaine, et une phase d'interprétation humaine est indispensable [Habert 96b]. Les auteurs pensent que leur méthode permet surtout d'amorcer la construction d'une ontologie, en exhibant les objets et relations susceptibles d'être représentés. Ils proposent d'ailleurs une étude comparative très intéressante avec les relations établies par des experts du domaine pour la construction d'une ontologie dans le domaine médical [Bouaud 97]. Pour les mêmes raisons que les travaux de Bourigault et al., la méthode proposée n'est pas directement exploitable pour rendre compte de manière synthétique du contenu des textes : elle demande une étude approfondie et itérative, peu compatible avec les besoins de l'analyse de l'information.

Les travaux d'Agarwal : classes sémantiques de noms et de verbes pour un domaine de spécialité

Agarwal propose une méthode semi-automatique pour construire des classes sémantiques qui soient utilisables dans le cadre du traitement automatique des langues naturelles

47. Les accents sont volontairement omis par les auteurs.

[Agarwal 94] [Agarwal 95]. Sa méthode s'applique sur des corpus de textes techniques. Elle est composée des étapes suivantes :

- un pré-traitement (étiquetage) puis une analyse syntaxique partielle qui met en évidence des relations de dépendances syntaxiques telles que sujet-verbe-objet, verbe-complément, nom-préposition-nom, ...,
- la définition d'un contexte pour chaque unité lexicale : pour un nom, on retient les verbes les plus fréquemment associés (en tant qu'objet et que sujet) ainsi que les prépositions,
- la classification à l'aide du système COBWEB/3[48] des termes en classes sémantiques, en utilisant la similarité des contextes,
- l'assignation d'une étiquette sémantique aux classes en faisant appel aux concepts de WordNet[49]. On obtient ainsi des motifs lexico-sémantiques, comme par exemple : *TREAT-VERB DISORDER with MEDICATION*
  où TREAT-VERB, DISORDER et MEDICATION sont des classes sémantiques regroupant plusieurs termes liés.

Le processus est itératif. Plusieurs classifications sont effectuées pour découvrir de nouveaux motifs lexico-sémantiques. L'assignation des étiquettes sémantiques est initialement manuelle. De même, chaque étape requiert une intervention manuelle pour corriger les motifs erronés.

La méthode proposée[50] est conçue pour l'acquisition lexicale : elle permet de repérer les motifs lexico-sémantiques qui seront ensuite recherchés dans les textes pour une analyse automatique. D'après [Pichon 97], les classifications obtenues sont pertinentes. Ce processus est simple à mettre en oeuvre, mais requiert un travail important de vérification et de correction à différentes étapes. Les résultats obtenus sont toutefois plus précis que les approches de Bourigault ou Habert : les motifs lexico-sémantiques sont proches des structures prédicat-arguments dont nous discutons dans la section 5.3.3.

Les travaux de Grivel et François : classes de termes pour analyser l'information bibliographique

Grivel et François proposent une station de travail pour analyser l'information bibliographique dans une perspective de veille scientifique [Grivel 95a] [Grivel 95b]. L'objectif est d'analyser le contenu d'un ensemble de documents à partir de leur résumés, titres, et descripteurs (termes) mais aussi d'analyser les acteurs, les institutions et les types de publications concernées. Pour cela, ils proposent une chaîne de traitement infométrique qui permet de construire des classes de termes et d´élaborer des cartes thématiques. Une carte thématique est définie comme « *une représentation de la topologie des relations entre des disciplines ou des thèmes de recherche, telles qu'elles sont matérialisées sous la forme de données bibliographiques* ».

48. Basé sur le même principe que COBWEB, présenté section 2.3.

49. Base lexicale, que nous avons présentée section 4.3.2.

50. Une approche itérative similaire est adoptée par Mikheev et Finch [Mikheev 95], qui propose des outils statistiques et linguistiques pour repérer des motifs lexico-sémantiques en utilisant la base lexicale WordNet.

Plus précisément, il s'agit d'utiliser trois types d'indicateurs différents pour répondre à la question : « qui fait quoi, avec qui et où? » [Grivel 97] :

- des termes, extraits des textes, comme « *indicateurs de la connaissance véhiculée par le document* »,
- des classes de termes, « *comme indicateurs des thèmes ou centres d'intérêt autour desquels s'agrègent l'information (articles, auteurs, institutions, périodiques)* »,
- une carte thématique, « *comme indicateur stratégique de la position relative des thèmes dans l'espace de connaissance couvert par les documents analysés* ».

L'extraction des termes et l'indexation (association des termes aux document correspondant) repose sur des outils externes, comme le système FASTER par exemple (cf. section 5.2.2). Grivel et François proposent deux outils différents pour constituer des classes de termes : SDOC et NEURODOC[51].

SDOC utilise la méthodes des mots associés : les termes sont regroupés selon leur co-occurrence dans une même phrase ou un même résumé. Plus précisément, la force de l'association entre deux termes est calculée par l'indice d'Equivalence, qui s'exprime par la formule :

$E_{ij} = C_{ij}^2/(C_i \times C_j)$
C_{ij} nombre de co-occurrences des termes i et j
C_i fréquence du terme i

Un algorithme de classification hiérarchique ascendante construit ensuite des groupes de termes n'excédant pas une taille maximale fixée par l'utilisateur. Par exemple, la figure 5.2 montre deux classes C_1 et C_2 contenant respectivement les termes T_1, ..., T_5 et T_6, ..., T_9. L'indice d'équivalence est précisé pour chaque lien. Lorsqu'une classe atteint le nombre maximal de termes autorisés, ce qui est le cas dans notre exemple pour C_1 ($max = 5$), les associations supplémentaires représentent des relations entre classes. Ainsi la relation entre les termes T_1 et T_6 est une association dite externe, liant les classes C_1 et C_2.

De manière similaire à l'indexation, un document est associés à une classe de termes

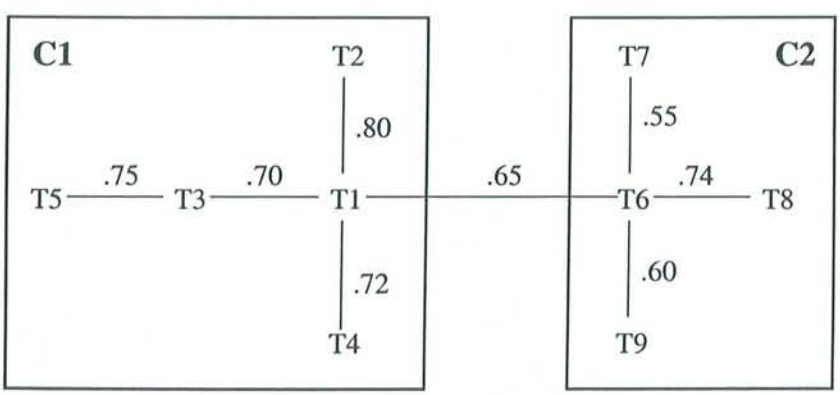

FIG. 5.2 – *Deux classes C_1 et C_2 de 5 termes maximum d'après [Grivel 95a]*

s'il possède dans sa liste de termes associée au moins un couple pouvant constituer une

51. Nous ne détaillons pas l'outil NEURODOC, qui est une approche complémentaire de celle utilisée par SDOC dont les principes sous-jacents sont similaires (constitution de classes et cartographie). NEURODOC est présenté dans [Grivel 97] et [Grivel 95a].

association interne ou externe de la classe. Par exemple, un document contenant les termes T_1 et T_3 est associé à la classe C_1, et un document contenant les termes T_1 et T_6 est associé aux classes C_1 et C_2. Les classes sont nommées par le terme de poids le plus fort parmi les termes de la classe.

Les classes sont ensuite visualisées sur une carte thématique. Deux indices permettent de positionner les classes sur une carte :

- la densité, qui représente la valeur moyenne des associations entre termes d'une classe ; intuitivement, plus cette valeur est forte, plus la classe est cohérente et structurée,
- la centralité, qui représente la valeur moyenne des associations entre les termes d'une classe et les termes d'autres classes (associations externes) ; plus cette valeur est forte, plus la classe est rattachée au réseau thématique.

La proximité de deux classes sur la carte indique seulement une similarité de structure, et non des contenus sémantiques liés. La carte permet « *à un utilisateur d'appréhender globalement et localement le contenu d'un corpus bibliographique* » [Grivel 95a].
La figure 5.3 est un exemple de carte construite à partir d'un corpus de textes scientifiques dans le domaine de l'agriculture. Sur la carte, la classe *DESHYDRATATION* est sélectionnée, et les classes qui lui sont associées sont encadrées : *SECHAGE*, *STRESS*, *SUCRES*. Les classes les plus significatives sont celles situées dans la partie haute à droite de la carte (par exemple, *SUCRES*, *VIGNE*, *MAIS*), correspondant à un densité et une centralité élevée. Les classes de faible densité et de faible centralité (par exemple, *TEXTURE*, *EAU*, *TERRE*) sont souvent peu pertinentes, car elles sont formées en dernier et sont très hétérogènes.

Les travaux de Grivel et François représente une première approche pertinente pour l'analyse de l'information. Ils se distinguent des travaux orientés vers l'acquisition (lexicale ou conceptuelle) par un souci de présentation et d'accessibilité des informations, au détriment d'une recherche en profondeur, itérative, qui n'est pas concevable pour une utilisation en veille scientifique ou technique.

Il est toutefois envisageable de pousser plus loin l'analyse proposée, en essayant de mieux caractériser les relations détectées par des moyens statistiques. Il faut alors s'intéresser aux approches symboliques qui permettent de repérer des structures qui rendent compte de façon moins fragmentée du contenu informatif des textes. Ce point est discuté dans la section suivante.

5.3.3 Les structures prédicat-arguments et les rôles thématiques pour une meilleure caractérisation des relations

Pour obtenir une meilleure caractérisation des relations entre les termes, il est nécessaire de s'intéresser à des fragments de phrases, notamment par le biais des structures prédicatives. Nous avons mentionné le travail de Harris sur les sous-langages, qui consiste à identifier au sein des phrases des opérateurs et leur arguments afin de mettre à jour le contenu informationnel des textes (section 5.2.1). La notion d'opérateur correspond à des actions, des états ou des événements qui s'appliquent sur des objets. Il en est de même pour le prédicat, terme qui possède un sens général, et qui est pourvu d'arguments qui précisent ce sens. Le prédicat et ses arguments forment une structure prédicative.

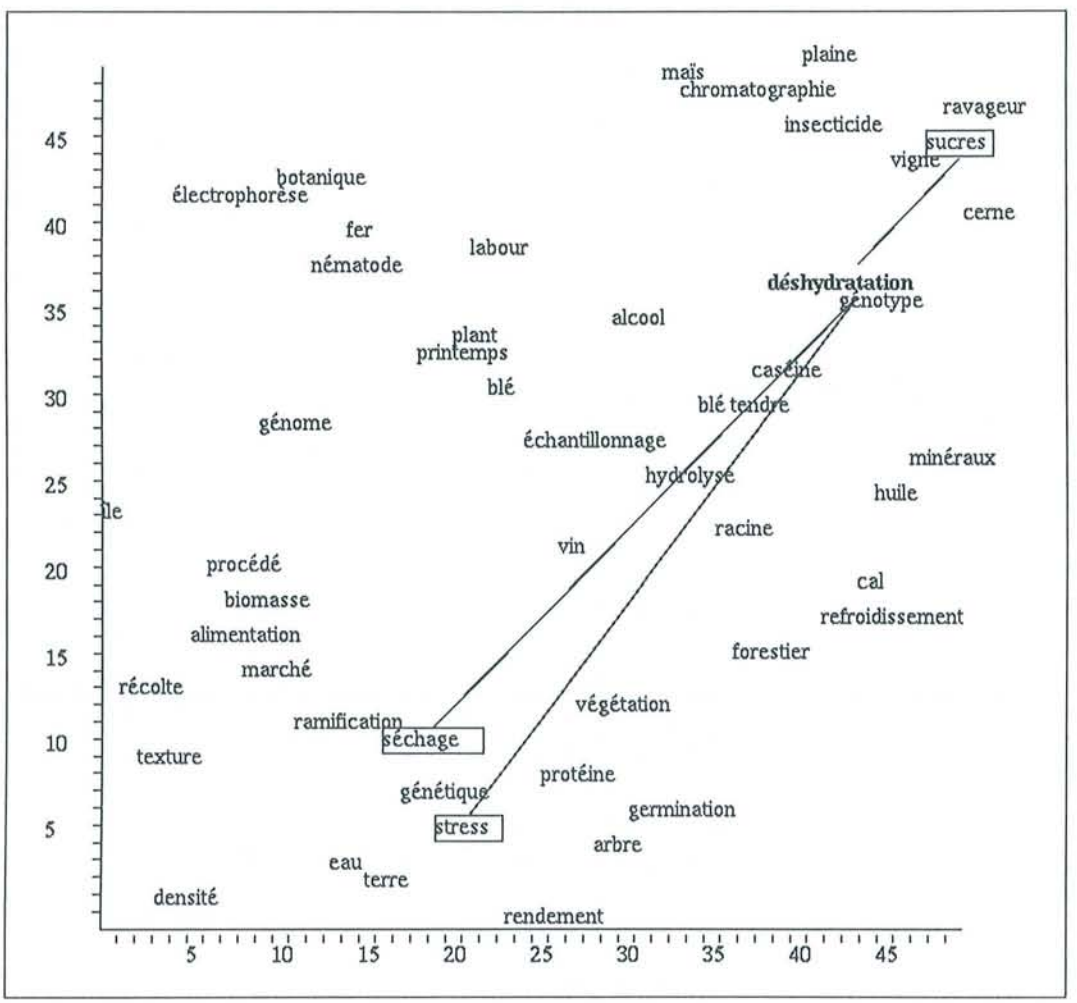

FIG. 5.3 – *Une carte thématique sur un corpus en agriculture, construite avec SDOC dans le cadre du projet ILC*

On associe à chaque prédicat une structure argumentale, qui spécifie le nombre d'arguments requis [Saint-Dizier 95]. Ces arguments représentent les éléments qui interviennent nécessairement dans l'événement ou l'état décrit par le prédicat. Il faut alors distinguer ces arguments essentiels des arguments optionnels, qui peuvent être utilisés pour compléter le sens d'un prédicat. Par exemple, pour le verbe *donner*, qui est un prédicat d'arité 3 (*X donne Y à Z*), il est possible de préciser où, quand, comment se déroule l'action. Ces compléments ne rentrent pas en compte dans la structure argumentale de *donner*. Notons que la distinction entre arguments essentiels et optionnels est loin d'être évidente, et qu'il n'existe pas de consensus pour déterminer de façon stricte l'arité d'un prédicat.
Les prédicats peuvent avoir différentes réalisation syntaxiques : si les verbes sont a priori tous des prédicats, les noms, adjectifs et prépositions peuvent avoir un comportement prédicatif. En particulier, les nominalisations (noms dérivés d'un verbe) ont souvent une structure argumentale réduite par rapport aux verbes, certains arguments devenant optionnels. Par exemple, *détruire* est un verbe d'arité 2 (*X détruit Y*), mais la nominalisation *destruction* ne requiert qu'un seul argument : *la destruction de la ville.*

Les arguments d'une structure argumentale n'entretiennent pas tous le même rapport avec le prédicat : si l'on dépasse le niveau strictement syntaxique (sujet, objet direct, objet indirect, ...), une certaine régularité sémantique est observable. Par exemple, de nombreux prédicats possèdent un argument qui fait intervenir un participant qui provoque ou est à l'origine de l'action correspondante. Cet argument est généralement considéré comme l'*agent* du prédicat. Les relations entre prédicat et arguments sont ainsi factorisées en un ensemble restreint de rôles thématiques, qui correspondent à des étiquettes sémantiques assignées aux arguments. Toutefois, il n'existe pas de définitions précises et consensuelles des rôles thématiques : d'une part, les ensembles de rôles, et leur sens, varient selon les auteurs ; d'autre part, il n'existe pas de véritables critères pour distinguer les rôles à assigner [Pugeault 95a].
Les rôles thématiques font le lien entre la syntaxe de la phrase et la sémantique. C'est pourquoi ils ont été utilisés en traitement automatique de la langue, malgré le manque de caractérisation et l'absence de méthode de reconnaissance. Les définitions de rôles les plus connues sont sans doutes celles de Fillmore [Fillmore 68], Jackendoff [Jackendoff 90] et Dowty[52]. Les rôles thématiques suivants sont parmi les plus utilisés, et possèdent une définition a peu près stable [Saint-Dizier 95] :

- *agent* : le participant désigné par le prédicat comme celui qui réalise ou cause l'action (premier argument de *manger*, *regarder*, *donner*),
- *patient* : le participant subissant l'action ou étant affecté par celle-ci (second argument de *tuer*, *manger*)
- *source* : le participant à partir duquel l'action se déroule (second argument de *partir*)
- *but* : la participant vers lequel l'action est dirigée, ou la motivation d'une action (second argument de *arriver*),
- *localisation* : la position, le lieu où l'action se déroule,
- *bénéficiaire* : le participant bénéficiant de l'action (second argument d'*offrir*)

Les rôles thématiques peuvent être hiérarchisés selon leur probabilité d'être pris comme sujet d'une phrase, ce qui est utile pour la conception de systèmes d'analyse, qui assignent des rôles aux entités d'une phrase. Ainsi, le rôle *agent* est considéré comme celui ayant le

52. Pour une étude critique, voir [Pugeault 95a].

plus de probabilité d'être en position de sujet : en cas d'ambiguïté entre deux rôles, *agent* est assigné. Enfin, il est possible de décomposer certains rôles en des fragments de sens, correspondant à des propriétés plus précises (proto-rôles). L'analyse est alors plus fine, et permet de traiter les cas où la continuité sémantique entre deux rôles rend le traitement difficile : c'est le cas des rôles *agent* et *patient*, dont la distinction est parfois subtile.

De manière générale, les chercheurs utilisant les rôles thématiques adoptent leur propre ensemble de rôles, qui sont adaptés à l'objectif visé et aux types de textes traités. On peut faire un parallèle avec les relations sémantiques utilisées en représentation des connaissances : il existe un ensemble restreint de relations incontournables, dont les définitions sont a peu près stables, et un ensemble potentiellement très grand de relations moins fréquentes et plus précises, dont les sens possèdent une certaine continuité et s'enchevêtrent.

De nombreux travaux se sont intéressés à l'automatisation de l'extraction de structures prédicatives dans un corpus de textes [Delisle 96]. Nous discutons de deux solutions récentes qui montrent les possibilités et les limites des systèmes d'extraction : le système PAPINS, conçu pour l'indexation sémantique de textes techniques de la société EDF, et le système RECIT, conçu pour la représentation sémantique de textes médicaux.

Le système PAPINS : extraction de structures prédicatives pour indexer les textes

Le système PAPINS (Prototype d'Analyse pour la Production d'INdex Sémantiques) a été conçu pour extraire des informations à partir de textes techniques, dans un but de synthèse et d'analyse [Pugeault 95a]. Plus précisément, il est appliqué à des textes relatant des actions de recherche et développement de la Direction des Etudes et Recherches d'EDF, qui sont des textes techniques complexes d'environ 300-400 mots. L'objectif est de pouvoir analyser automatiquement l'activité interne d'EDF, en montrant qui fait quoi, et quels sont les résultats disponibles. Il est donc très proche de l'analyse de l'information, telle que nous l'avons définie.

Pugeault a opté pour une approche linguistique, utilisant la sémantique lexicale et mettant en oeuvre de nombreuses connaissances syntaxiques et sémantiques sur le domaine considéré. L'idée est d'identifier et de représenter trois types de connaissances : « *les actions caractérisant des états ou des événements, les objets intervenant dans le procès d'une action, et les relations existants entre une action et les objets qui lui sont associés* ». Pugeault a donc choisi d'exploiter des structures prédicatives, extraites à partir des phrases des textes du corpus.

PAPINS est composé de 3 niveaux distincts :

- un niveau pragmatique (niveau 1) qui à partir de textes pré-traités identifie des phrases et des fragments de phrases, et les classe en quatre articulations : thème, motivations, problèmes, réalisations. Le classement est opéré par des règles d'extractions qui repèrent des marqueurs linguistiques dont la liste est établie au préalable,
- un niveau linguistique (niveau 2) qui extrait automatiquement des structures prédicatives à partir des fragments de phrases. Cette étape requiert un analyseur morpho-syntaxique et de nombreuses connaissances linguistico-sémantiques (détaillées ci-après),

- un niveau conceptuel (niveau 3) qui utilise le formalisme des structures lexico-conceptuelles de Jackendoff[53] [Jackendoff 90] pour « *représenter les formes prédicatives sous un format plus générique* », et permettre par exemple de travailler sur plusieurs langues.

Le niveau pragmatique est intéressant mais requiert des textes ayant toute la même typologie et la même structure, ce qui n'est pas le cas en général. Nous ne développerons donc pas cet aspect. Le niveau linguistique montre que l'extraction de structures prédicatives à partir de phrases complexes est possible.

La solution proposée par Pugeault est d'utiliser une grammaire partielle avec un ensemble de règles d'assignations de rôles thématiques, qui prennent en compte la syntaxe et la sémantique des prédicats et des arguments mis en jeu dans la phrase. Elle définit un ensemble de rôles thématiques, inspirés des travaux de Levin et Jackendoff principalement. Elle montre qu'il faut disposer initialement :

- d'une liste des prédicats, qui peuvent être des verbes, des noms, et des prépositions,
- d'une organisation des prédicats en classes sémantiques, qui regroupent des prédicats de sens voisins. A chaque classe est attribuée une grille thématique, qui donne les rôles associés aux arguments. Par exemple, pour le prédicat *manger*, la grille thématique est *(agent, thème)* (pour la phrase*Julien mange des fruits*) ou *agent* (pour la phrase *Julien mange*),
- d'une organisation des prépositions (*avec*, *pour*, *dans*, ...) en classes, dont certaines sont prédicatives et possèdent une grille thématiques, et d'autres sont grammaticales et ne jouent pas le rôle de prédicats,
- d'une organisation hiérarchique des arguments des prédicats, afin de pouvoir exprimer des restrictions de sélections,[54]
- des règles d'assignation de rôles thématiques, qui assignent un rôle r étant donné un ensemble de contraintes : l'argument doit être de catégorie syntaxique N, de type X, le prédicat doit être de catégorie C, de classe sémantique S, la préposition (éventuelle) de type P. Par exemple, une règle assigne le rôle *agent effectif* si le prédicat est un verbe ou un nom appartenant à la liste [*caractérisation, réalisation, ..., réponse, séparation*] et si l'argument est un nom commun ou nom propre de type « humain ».

Pugeault a ainsi défini environ 90 règles d'assignation de rôles thématiques pour son prototype. Elle considère que celui-ci conduit à 78,5% de représentations prédicatives correctes. Les problèmes rencontrés sont liés principalement à [Pugeault 95b] :

- l'analyseur morpho-syntaxique (étiquetage lexical),
- l'incomplétude du lexique, qui contient les informations syntaxiques et sémantiques pour chaque unité lexicale,
- l'incomplétude de la grammaire utilisée : les phénomènes complexes tels que l'anaphore, les références, les ambiguïtés lexicales, ... ne sont pas traités.

Le niveau conceptuel (niveau 3) n'a pas fait l'objet d'une implémentation et n'a pas été exploité. Par contre, Pugeault s'est intéressée par la suite à la génération automatique de

53. Lexical Conceptual Structure, LCS.

54. Les restrictions de sélection sont des contraintes sémantiques sur les arguments d'un prédicat, utilisées pour filtrer les arguments valides et non valides pour un prédicat. Dans le cas de PAPINS, une restriction de sélection est soit un type sémantique, soit une combinaison de disjonctions et de conjonctions sur des types.

synthèse de textes techniques à partir de la représentation prédicative du corpus [Pugeault 96]. Il s'agit de pouvoir répondre automatiquement à des questions posées par un utilisateur, en s'appuyant sur la connaissance extraite et disponible sous forme de structures prédicatives. Un algorithme en trois étapes permet alors : 1) d'apparier la question à un type de question qui permet de guider vers un schéma de construction de la réponse ; 2) de rechercher les structures prédicatives répondant à la question ; 3) de générer des phrases en langages naturel à partir des structures prédicatives sélectionnées.

Le travail de Pugeault montre que l'extraction automatique de structures prédicatives est possible et permet de cerner les principales difficultés posées : le besoin important de connaissances sur les unités lexicales (prédicats et arguments) et la complexité des phrases à traiter. Toutefois, le prototype PAPINS ne permet de traiter qu'un ensemble très limité de textes, dans un domaine restreint. L'extension à un autre domaine requiert un nouveau travail de collecte d'informations lexicales et sémantiques qui représente une charge importante de travail. Cette approche est sans doute trop ambitieuse pour une tâche d'analyse de l'information : il n'y a pas vraiment de sélection ou de synthèse de l'information, mais une formalisation du contenu sous forme prédicative, sorte de réécriture normalisée des textes. L'analyse sous forme de questions-réponses, proposée avec la génération de phrases en langues naturelles, ne permet pas de cerner l'ensemble des informations d'un corpus de textes : elle est plus proche de la recherche d'information, et suppose la formulation d'une requête. Nous pensons que l'objectif de l'analyse de l'information doit conduire à des systèmes qui effectuent une analyse moins complète et plus sélective, pour offrir une vue plus synthétique du corpus.

Le système RECIT : représentation du contenu informationnel de textes médicaux

Le système RECIT (REprésentation du Contenu Informationnel des Textes médicaux) a été conçu pour extraire automatiquement l'information pertinente de textes médicaux, afin de la stocker et de l'utiliser directement pour la recherche d'information [Rassinoux 94]. L'approche choisie se veut pragmatique, et repose sur une description approfondie de la sémantique du domaine traité, la chirurgie digestive, et la prise en compte des spécificités des textes analysés, des lettres de sorties d'une clinique.

Le système RECIT effectue une analyse des textes en deux phases :

- une décomposition des phrases en fragments significatifs, par un *traitement des proximités* qui prend en compte les proximités syntaxiques et sémantiques des mots pour les associer en structures significatives,
- une construction d'une représentation canonique des phrases à l'aide du formalisme des graphes conceptuels.

Rassinoux sépare clairement deux niveaux, lexical et conceptuel :

- le niveau lexical concerne les unités lexicales, c'est-à-dire les mots et les expressions idiomatiques (groupes de mots considérés comme une séquence non décomposable), et leur propriétés linguistiques,
- le niveau conceptuel concerne la description des connaissances sur le domaine, exprimées par des concepts et relations entre concepts, et faisant appel au formalisme des graphes conceptuels.

Les concepts sont organisés hiérarchiquement, et divisés en plusieurs catégories : acteurs, événements, attributs, valeurs et modalités. Les relations sont nombreuses et réparties en quatre types permettant de prendre en compte les différents niveaux d'informations : relations modales (négation, possibilité, ...), relations thématiques, relations temporelles (durée, temps, ...) et relations inter-phrases (but, cause, condition, ...).

Tout le traitement du système RECIT consiste à partir des mots et des expressions idiomatiques[55], à les associer, et à calculer une représentation utilisant les concepts et relations définies sur le domaine. La représentation conceptuelle est en fait très similaire aux structures prédicatives, puisqu'elle fait largement appel aux relations thématiques. Elle est cependant plus puissante, car elle permet également de représenter des relations non prises en compte dans les structures prédicatives, telles que les modalités ou les relations entre phrases.

Tout comme le système de Pugeault, RECIT nécessite de nombreuses connaissances. Chaque unité lexicale est décrite par ses propriétés linguistiques et renvoie à une description conceptuelle (souvent limitée à un concept). Par exemple, les mots *fièvre*, *afébrile* et *cholécystectomie* sont décrits de la façon suivante dans le dictionnaire[56] :

fièvre : nom(féminin, singulier), [cl_symptome]
afébrile : adjectif(_, singulier), [NEG(cl_symptome : fievre)]
cholécystectomie : nom(féminin, singulier),
[cl_trait_chirurgical : ablation, [THEME(cl_organe : vesicule_biliaire)]]

La première information renseigne sur la catégorie syntaxique et les traits flexionnels du mot. La deuxième information associe une description conceptuelle au mot : ainsi, la cholécystectomie est décrite comme étant une ablation de la vésicule biliaire.

La figure 5.4 illustre les différentes étapes du traitement des proximités et les connaissances nécessaires. En plus du dictionnaire, quatre types de règles sont utilisées :

- des règles morphologiques pour reconnaître certains mots absents du dictionnaire et dérivables à partir d'autres mots,
- des règles d'associations fréquentes, qui associent des mots voisins. Par exemple, une règle associe deux mots X et Y dont les concepts sont *cl_quantité* et *cl_unité_temps* en une séquence XY dont le concept est *cl_duree*,
- des règles de résolution des ambiguïtés syntaxiques, qui permettent de déterminer la catégorie syntaxique d'un mot lorsqu'il y a plusieurs possibilités, en se basant sur les catégories syntaxiques des mots voisins,
- des règles de compatibilité syntaxico-sémantique, qui étant donné deux mots dont les concepts sont X et Y et une structure syntaxique à respecter, établit une relation sémantique entre X et Y. Par exemple, si X est un nom qui correspond à un des concepts *cl_zone_corps*, *cl_organe* ou *cl_membre*, et si Y est un adjectif correspondant au concept *cl_region*, alors la séquence XY se voit associer une relation PARTIE_DE entre les concepts de X et Y,

55. Les expressions idiomatiques, telles que les définit Rassinoux, sont en fait des groupes nominaux, la plupart du temps assimilables à des termes.

56. Par souci de clarté, nous simplifions la notation utilisée par Rassinoux. Les classes sémantiques (cu concepts), par convention, ont un nom commençant par « cl_ ». Les relations sémantiques sont en caractères majuscules.

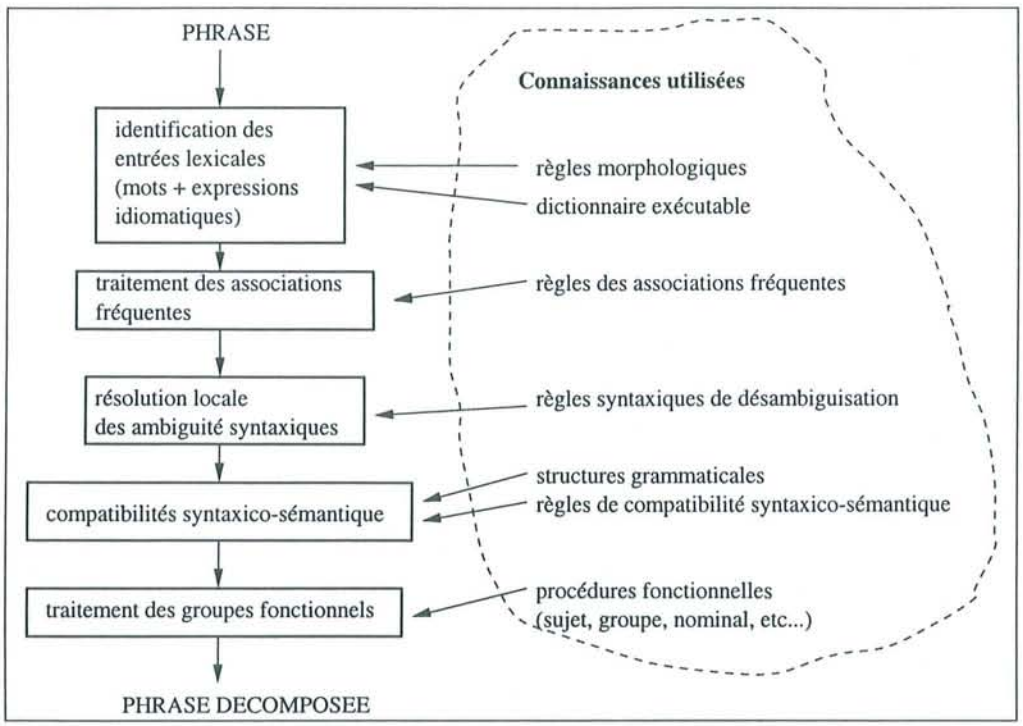

FIG. 5.4 – *Les étapes du traitement des proximités du système RECIT d'après [Rassinoux 94]*

Le traitement des proximités permet d'aboutir à des phrases décomposées, dont les principales unités significatives ont été établies. La deuxième étape effectue la construction des graphes conceptuels, en s'appuyant sur trois types d'informations :

- des schémas conceptuels, qui associent à chaque concept un ensemble de relations susceptibles d'être reconnues, et optionnellement une liste de concepts susceptibles d'être mis en relation. Par exemple, au concept *cl_douleur* est associé la relation LOCALISATION, au concept *cl_signe_symptome* est associé (entre autres) la relation CAUSE et le concept *cl_maladie*,
- des descriptions syntaxico-sémantiques des verbes, qui associent à chaque verbe un concept ainsi que des relations conceptuelles et la façon dont elles sont réalisées grammaticalement par des groupes fonctionnels (sujet, complément, ...). Ainsi, pour le verbe *hospitaliser*, le concept associé est *cl_trait_general :hospitalisation*, la relation INSTIGATEUR est réalisée par la fonction grammaticale *sujet* et la relation EXPERIENT[57] est réalisée par la fonction grammaticale *complément d'objet direct*,
- des connaissances par défaut permettent de compléter la description des relations. Par exemple, la relation LOCALISATION se voit associer par défaut le concept *cl_partie_corps*.

Les trois types d'informations se combinent, notamment par l'utilisation d'un mécanisme d'héritage, et permettent ainsi de caractériser complètement les différents éléments des phrases. C'est la description des verbes qui permet de faire le lien entre la syntaxe (groupes fonctionnels) et la sémantique (concepts et relations sémantiques). Le résultat final est un

57. Désigne celui qui subit une action.

graphe conceptuel qui formalise le contenu informationnel de la phrase analysée. La figure 5.5 illustre le graphe obtenu pour une phrase du corpus médical.

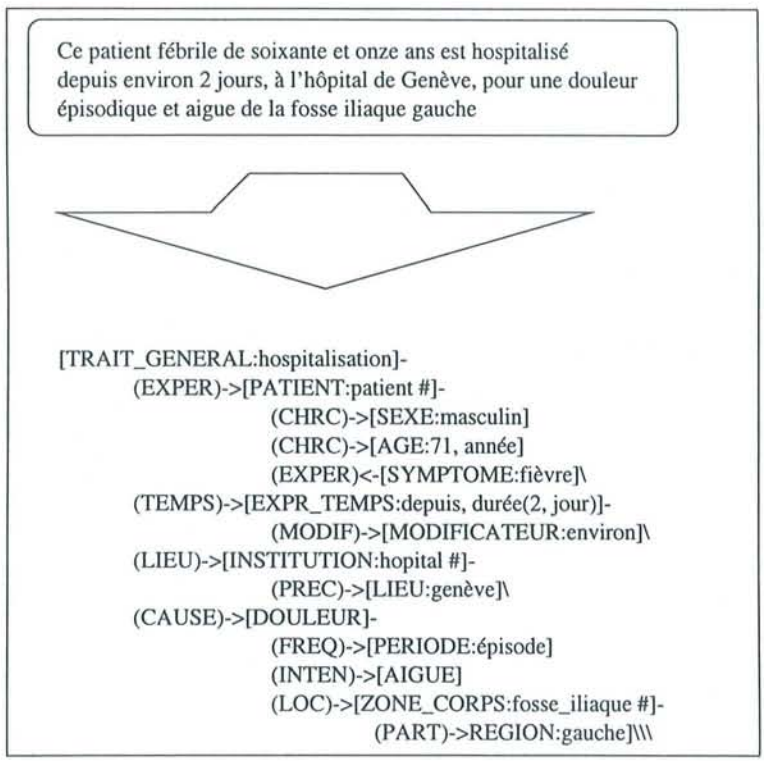

FIG. 5.5 – *Graphe conceptuel obtenu à partir d'une phrase du corpus d'après [Rassinoux 94]*

Le système RECIT, comme le système PAPINS, n'effectue aucune synthèse de l'information : l'exploitation de l'information extraite est réalisée par l'intermédiaire de requêtes. Il s'agit en fait de la recherche d'information conceptuelle décrite par Mauldin (cf. section 5.1). RECIT effectue toutefois une analyse complète de textes, qui va plus loin que la simple représentation de structures prédicatives : des inférences sont réalisées, le texte est analysé en profondeur. Enfin RECIT est adapté pour prendre en compte plusieurs langues, le niveau conceptuel étant indépendant de la langue.

Conclusion

Les structures prédicatives, ou des représentations sémantiques similaires comme dans RECIT, permettent de mieux appréhender le contenu informationnel. Des solutions existent pour les extraire directement des textes comme nous l'avons vu avec le système RECIT et le système PAPINS. Mais elles demandent beaucoup de connaissances à la fois syntaxiques et sémantiques : elles ne sont donc pas applicables hors d'un domaine réduit, et les résultats ne sont pas suffisamment satisfaisants pour pouvoir être utilisés directement, sans intervention humaine.
De plus, l'extraction de chacune des phrases d'un texte pour en déduire une représentation normalisée ne conduit pas à une réduction des informations, mais seulement à une

tentative de formalisation. Pour répondre à nos besoins en analyse de l'information, il faut proposer une analyse plus simple. Il est possible de laisser de côté certaines structures trop complexes, et de se focaliser sur les autres : les structures verbales peuvent par exemple être négligées au profit des structures nominales, beaucoup plus fréquentes dans les textes techniques, comme nous le verrons plus loin. Nous montrons dans la section suivante comment l'extraction de structures prédicatives peut être mise à profit dans une chaîne de traitement d'analyse de l'information.

5.4 Vers une chaîne de traitement complète pour l'analyse de l'information

Nous avons décrit plus haut les travaux de Grivel et François, qui proposent une approche d'analyse de l'information utilisant des méthodes infométriques (section 5.3.2). Cette approche a été poursuivie dans le cadre du projet ILIAD[58], dont l'objectif a été la construction d'une chaîne de traitement automatique pour l'analyse de l'information contenue dans des corpus de grande taille [Toussaint 98] [Toussaint 96]. L'expérimentation d'ILIAD a été réalisée sur le domaine de l'agriculture sur un corpus de textes en français de 2,5 Mb (résumés de notices bibliographiques).

Cette chaîne repose sur un ensemble d'outils existants, déjà présentés : ACABIT, FAS-

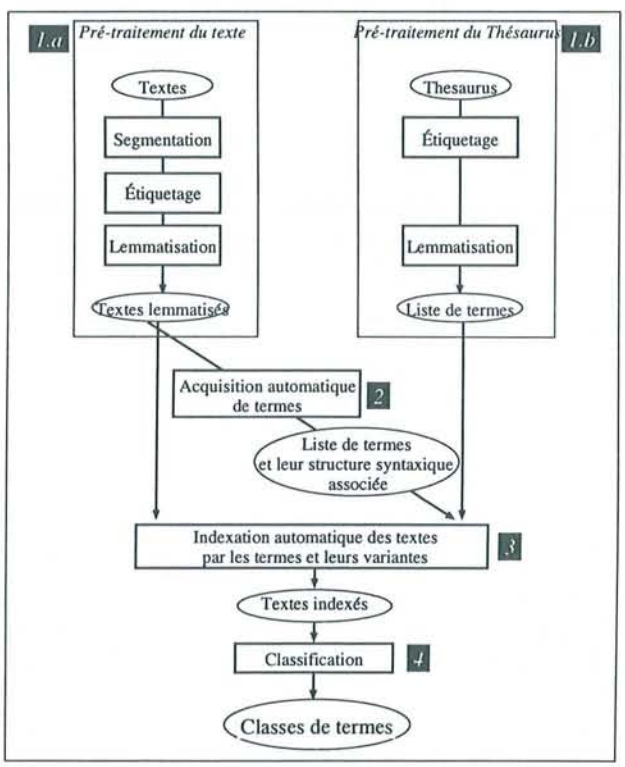

FIG. 5.6 – *Architecture de la chaîne de traitement ILIAD d'après [Toussaint 98]*

58. Informatique Linguistique et Infométrie pour l'Analyse de grands fonds Documentaires, GIS Sciences de la Cognition.

TER et SDOC. Nous en donnons son architecture figure 5.6. L'analyse des textes nécessite l'utilisation d'un thésaurus qui fournit une liste de termes initiale sur le domaine.
Les étapes (1.a) et (1.b) consistent en un pré-traitement des textes et du thésaurus, nécessaires pour l'utilisation des outils d'extraction.
L'étape (2), optionnelle, est réalisée par ACABIT, et permet d'enrichir le vocabulaire des termes utilisés.
L'étape (3) est réalisée par FASTER et permet d'indexer les textes par les termes et leur variantes.
Enfin, l'étape (4), qui construit les classes de termes, est réalisée par SDOC. La figure 5.7

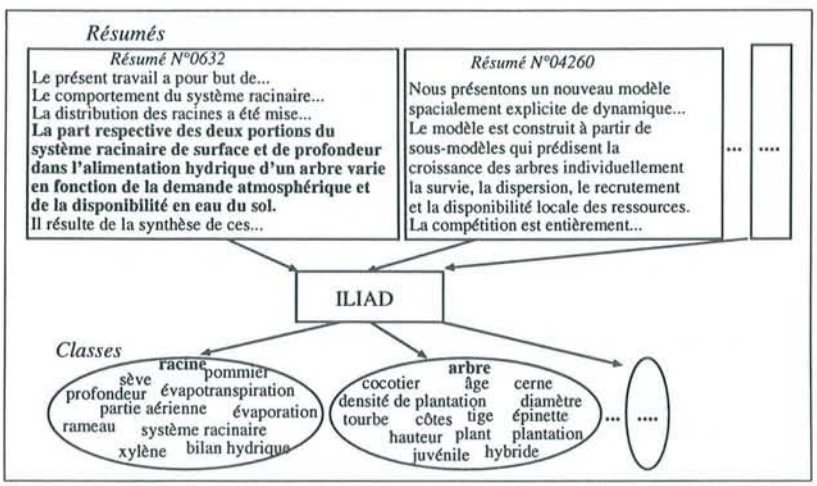

FIG. 5.7 – *Textes initiaux et classes de termes obtenues, d'après [Toussaint 98]*

illustre le traitement effectué par la chaîne ILIAD. Les classes de termes sont exploitées par les outils de visualisation décrits avec SDOC (voir section 5.3.2).
Cette chaîne est une association originale de techniques de traitement automatique de la langue et d'infométrie, utilisée par des experts pour l'analyse de l'information en agriculture et en médecine.

Les informations fournies par les classes de termes restent toutefois assez imprécises : on connaît les associations entre termes mais non la nature de ces relations. Il est donc nécessaire, pour obtenir plus de renseignements sur le contexte d'une association, de consulter les documents où elle apparaît. Pour éviter un tel retour aux documents, qui demande souvent beaucoup de temps, il est possible de mettre en oeuvre des solutions plus fines, telle que l'extraction de structures prédicatives.

Nous proposons ainsi d'utiliser deux moyens supplémentaires pour compléter l'analyse [Capponi 97b] :

- une organisation hiérarchique sur les termes, qui permet de structurer les classes de termes et d'identifier les associations de type paradigmatique (relations de spécialisation) entre termes,
- une extraction des structures prédicatives mettant en jeu au moins un terme d'une classe, afin d'identifier la nature de l'association de type syntagmatique entre ces termes, et de proposer un contenu informationnel plus précis.

Ces deux moyens permettent un accès plus fin au contexte d'une association et permettent

d'éviter un retour aux documents, comme nous le montrons dans le chapitre suivant. Cependant, l'extraction des structures prédicatives conduit à une masse d'information beaucoup plus conséquente. Pour limiter l'extraction, nous ne considérons que certaines structures syntaxiques, les structures nominales, et uniquement les phrases mettant en jeu les termes d'une classe. Pour présenter ces informations de façon synthétique, nous avons recours à notre processus de généralisation de structures prédicatives : le contenu informatif est ainsi présenté en un ensemble réduit de généralisations.

Nous détaillons ces solutions dans le cadre d'une expérimentation en analyse de l'information sur un corpus de l'agriculture, présentée au chapitre suivant.

6

Les structures prédicatives et leurs généralisations pour l'analyse de l'information

Nous avons proposé deux moyens de caractériser de façon plus précise une classe de termes : (1) une organisation hiérarchique des termes ; (2) une extraction de structures prédicatives mettant en jeu les termes de la classe. Dans ce chapitre, nous discutons d'une expérimentation en analyse de l'information mettant à profit les structures prédicatives et leurs généralisations. Nous montrons tout d'abord comment l'utilisation d'un thésaurus permet d'obtenir une hiérarchie de termes (section 6.2), puis comment les structures prédicatives sont utilisées pour identifier les relations de type syntagmatique et les synthétiser (section 6.3). Nous détaillons ensuite notre prototype mettant en oeuvre le processus de généralisation de structures prédicatives et son interface (section 6.4). Nous pouvons alors présenter l'évaluation de la généralisation pour l'analyse de l'information au travers d'une expérimentation avec un documentaliste expert (section 6.5).

6.1 Le domaine d'application : résumés bibliographiques sur l'agriculture

Nous utilisons pour notre expérimentation un corpus de textes issus d'un fond documentaire dans le domaine de l'agriculture. Ces textes sont des résumés en français, extraits de la base PASCAL[59]. Le corpus utilisé comprend 2069 résumés provenant d'articles de plusieurs revues scientifiques et portant sur divers thèmes. Ces textes ont été analysés par la chaîne de traitement ILIAD présentée au chapitre précédent, et ont conduit à la création de 50 classes de termes. La liste initiale de termes est fournie par le thésaurus AGROVOC[60], comprenant environ 15 000 termes.

La figure 6.1 représente un texte issu de notre corpus, dont les termes ont été soulignés. Nous nous focalisons tout au long de ce chapitre sur une des classes générées par SDOC

59. PASCAL est une base documentaire scientifique développée et maintenue par l'INIST-CNRS.

60. Thésaurus multilingue développé par AGRIS (Organisation des Nations Unies pour l'Alimentation et l'Agriculture, unité de traitement AGRIS).

TITRE :
Identification et dosage des amines biogènes dans les farines d'origine animale
RÉSUMÉ :
L'identification et le dosage des amines par formation de dérivés orthophtalaldéhyde puis passage en chromatographie en phase liquide ont été appliqués à la détermination des teneurs en cadavérine, histamine, phényléthylamine, putrescine, spermidine, tryptamine et tyramine dans les farines de viande et de poisson destinées à l'alimentation animale. Les répétabilités, les taux de récupération et les limites de quantification de la méthode ont été étudiés. Trente-sept échantillons de provenance connue ont été analysés. Les résultats obtenus montrent les différences de composition en amines entre farine de poisson et farine de viande. La composition en amines des farines de poisson est différente en fonction de l'origine géographique de ces farines.

FIG. 6.1 – *Un résumé extrait du corpus, dont les termes sont soulignés*

pour illustrer le traitement proposé. Nous avons choisi la classe CHROMATOGRAPHIE, qui possède une centralité et une densité élevée[61], assurant ainsi que la classe est cohérente et non isolée par rapport aux autres classes.

Rappelons qu'une classe est un ensemble de termes reliés par des liens de co-occurrence. La classe CHROMATOGRAPHIE contient 12 termes et 19 liens de co-occurrence.
Les termes de CHROMATOGRAPHIE sont donnés par la table 6.1, qui précise leur fréquence et leur poids. La table 6.2 donne les liens de co-occurrence entre les termes, avec leur poids et leur nombre d'apparition dans le corpus. L'hypothèse qui est à la base de SDOC, à savoir que les associations calculées entre termes sur la base de leur co-occurrence ont une interprétation sémantique, est vérifiée par le fait que les experts sont capables de verbaliser les classes. Ainsi, la classe CHROMATOGRAPHIE est centrée autour de la technique de *chromatographie*, qui est utilisée pour analyser des substances chimiques variées. La classe contient de plus les opérations pour lesquelles la technique est utilisée (*dosage* et *purification*), les substances qui sont analysées (qui sont toutes des *amines*), et certains produits de l'agriculture dans lesquels on peut trouver ces substances (*produit laitier*, *miel*, *boisson*).

61. Ces notions sont définies section 5.3.2, dans le paragraphe décrivant SDOC.

Terme	Fréquence	Poids
chromatographie	36	0.31
putrescine	9	0.24
amine biogène	8	0.28
histamine	6	0.14
boisson	17	0.14
polyamine	6	0.10
spermidine	4	0.10
spermine	3	0.10
dosage	41	0.10
produit laitier	5	0.03
purification	11	0.03
miel	7	0.03

TAB. 6.1 – *Termes de la classe* CHROMATOGRAPHIE *avec leur fréquence et poids*

ID	Association	P	C
L1	spermidine & spermine	0.75	3
L2	amine biogène& histamine	0.52	5
L3	polyamine & spermine	0.50	3
L4	polyamine & putrescine	0.46	5
L5	putrescine & spermidine	0.44	4
L6	polyamine & spermidine	0.38	3
L7	putrescine & spermine	0.33	3
L8	histamine & putrescine	0.30	4
L9	amine biogène & putrescine	0.22	4
L10	amine biogène & chromatographie	0.12	6
L11	amine biogène & boisson	0.07	3
L12	chromatographie & dosage	0.06	9
L13	chromatographie & produit laitier	0.05	3
L14	chromatographie & putrescine	0.05	4
L15	amine biogène & dosage	0.05	4
L16	chromatographie & histamine	0.04	3
L17	chromatographie & purification	0.04	4
L18	chromatographie & miel	0.04	4
L19	boisson & chromatographie	0.01	3

TAB. 6.2 – *Associations de la classe* CHROMATOGRAPHIE *avec leur poids (P) et leur nombre de co-occurrences (C)*

6.2 Réutiliser le thésaurus AGROVOC pour organiser les termes en hiérarchie

La classe de termes en elle-même constitue un ensemble d'informations non structurées. L'organisation hiérarchique permet de mettre en évidence les différentes catégories de concepts mises en jeu par les termes : événements, états, entités vivantes, artefacts, propriétés, ... La construction d'une hiérarchie est une tâche longue et difficile. Pour limiter le coût engendré, il semble pertinent de réutiliser une source de connaissances. Nous avons exploité le thésaurus AGROVOC, utilisé pour constituer la liste initiale de termes traitée par FASTER (cf. section 5.2.2).

Un thésaurus est un dictionnaire de termes, où chaque entrée possède des liens avec d'autres entrées. Étant donné un terme t, ces liens sont de quatre types :

- termes génériques, possédant un sens plus général que le terme t,
- termes spécifiques, possédant un sens plus restreint que le terme t,
- termes synonymes, possédant un sens voisin du terme t,
- termes liés, possédant un lien sémantique non précisé avec le terme t.

AGROVOC est constitué d'environ 15 000 termes principaux possédant une entrée dans le thésaurus, et d'environ 7 000 termes synonymes qui n'ont pas d'entrées et sont liés à un terme principal unique. La figure 6.2 donne deux entrées extraites d'AGROVOC. Les deux

entrée : AMINE
termes génériques : COMPOSÉ AMINÉ
termes spécifiques : AMINE BIOGÈNE, HYDROXYLAMINE, POLYAMINE

entrée : PRODUIT DE LA RUCHE
terme synonyme : EXTRAIT DE RAYON
termes génériques : PRODUIT ANIMAL
termes spécifiques : CIRE D'ABEILLE, GELÉE ROYALE, MIEL, PROPOLIS
termes liés : PRODUCTION DE MIEL

FIG. 6.2 – *Deux entrées du thésaurus* AGROVOC, AMINE *et* PRODUIT DE LA RUCHE

premiers types de liens du thésaurus (génériques et spécifiques) structurent les termes en hiérarchie. Les termes n'étant pas tous connexes, il existe de fait une multitude de hiérarchies de tailles diverses (environ 1400 hiérarchies). Il est donc nécessaire de compléter le thésaurus pour uniformiser la structure et obtenir une hiérarchie unique.

La modification du thésaurus consiste à ajouter une hiérarchie de catégories abstraites, qui connecte les différentes parties du thésaurus. Par exemple, l'ajout d'une catégorie abstraite *produit* permet d'aboutir à la hiérarchie illustrée figure 6.3. La réutilisation d'un thésaurus peut poser quelques problèmes dans le mesure où c'est un outil principalement destiné à une utilisation humaine. Par exemple, la classification n'est pas homogène : comme l'illustre la hiérarchie de la figure 6.3, les différents points de vue sur un concept (e.g., classement selon *l'origine* ou la *fonction* d'un produit) sont traités indifféremment. De même,

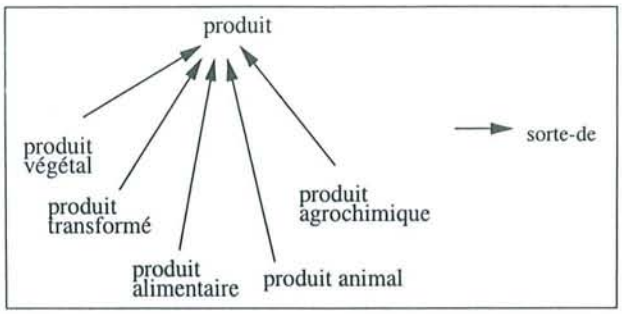

FIG. 6.3 – *Structuration des hiérarchies du thésaurus par ajout de la catégorie abstraite produit*

certains termes du thésaurus sont regroupés selon un sous-domaine plutôt que selon les propriétés intrinsèques des concepts correspondants. Par exemple, la hiérarchie du terme *aquaculture* contient un « objet physique » (*étang*) parmi un ensemble d'événements (figure 6.4)[62]. Ces problèmes se posent naturellement à tous les concepteurs de hiérarchies, que ce

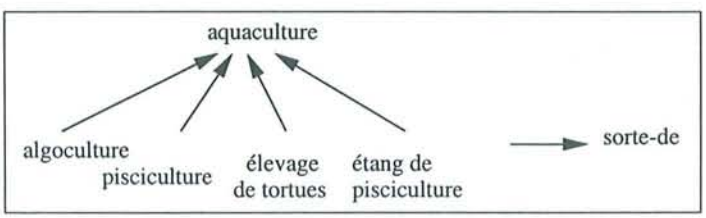

FIG. 6.4 – *La classification du thésaurus n'est pas toujours homogène*

soit pour construire un thésaurus ou une ontologie, pour le traitement du langage naturel ou pour les systèmes à base de connaissances. Il est généralement admis que ce travail ne peut être qu'empirique et restreint à un domaine précis ou à une tâche précise [Bachimont 95]. Dans la mesure où il s'agit d'une classification sémantique, il semble impossible d'obtenir une hiérarchie totalement satisfaisante. Il existe trop de ramifications entre les sens des différentes notions possibles pour pouvoir les cerner à l'aide d'une simple hiérarchie. Les réflexions les plus abouties sur cette question sont sans doutes celles de Zweigenbaum et al., qui préconisent de se restreindre à des arbres (hiérarchie strictes) et proposent des critères de regroupement des notions [Zweigenbaum 94] [Charlet 94] [Bachimont 95]. Dans le cadre de l'analyse de l'information, la structure idéale est celle construite par ou avec l'utilisateur, afin qu'il puisse exploiter et interpréter au mieux la classification.

La structure hiérarchique établie sur les termes permet de structurer les classes établies par SDOC. Nous montrons sur la figure 6.5 la projection de la classe CHROMATOGRAPHIE sur la hiérarchie. Les termes de la classe sont encadrés. La visualisation graphique de la classe met en évidence les différents groupes de termes présents au sein de la classe. Deux concepts principaux sont clairement identifiés : les *amines*, qui sont des substances, et les *produits*, qui sont les objets issues de l'agriculture. A côté de ces objets, on trouve trois processus : *chromatographie*, *dosage* et *purification*.

La hiérarchie permet d'identifier, parmi les associations de la classe de termes, celles

62. De plus, le lien entre *aquaculture* et *étang* est difficilement interprétable comme « sorte-de ».

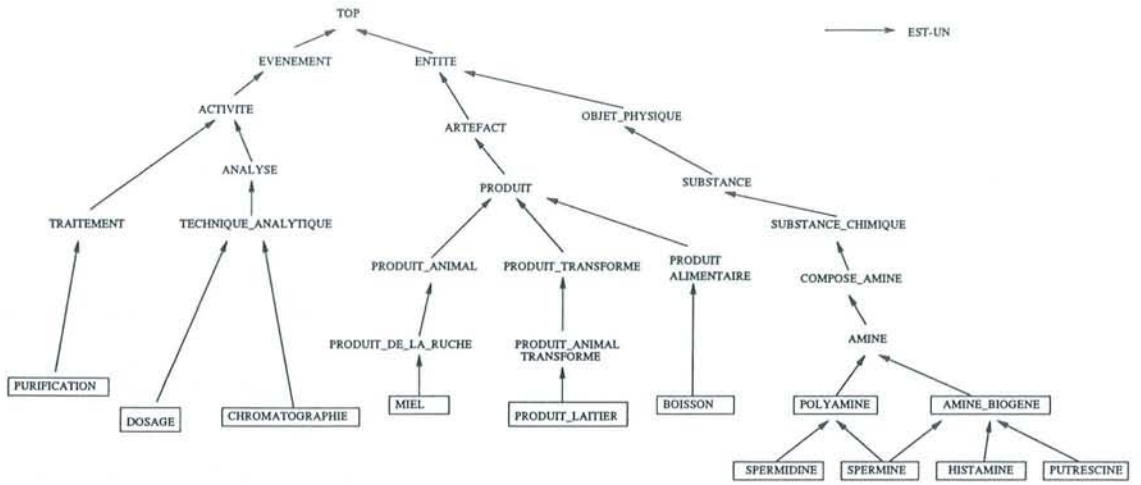

FIG. 6.5 – *Termes de la classe* CHROMATOGRAPHIE *replacés dans la hiérarchie*

qui correspondent à une relation paradigmatique. Les liens L_1 à L_9 de la classe CHROMATOGRAPHIE sont des liens paradigmatiques. L'importance numérique de ces liens (9 sur 19 pour la classe CHROMATOGRAPHIE) montre que ce ne sont pas les seules relations syntagmatiques qui sont exhibées par les méthodes infométriques. Nous nous sommes limités aux relations de généralité, mais il serait envisageable d'introduire d'autres relations paradigmatiques, telles que les relations partie-tout [Winston 87]. Cela nécessiterait alors de les introduire et de les représenter dans la hiérarchie des termes.

La figure 6.6 illustre les termes de la classe CHROMATOGRAPHIE structurés par la hiérarchie ainsi que les liens de co-occurrences qui ne correspondent pas à des relations paradigmatiques. Comment rendre compte de ces liens ? C'est l'extraction de structures

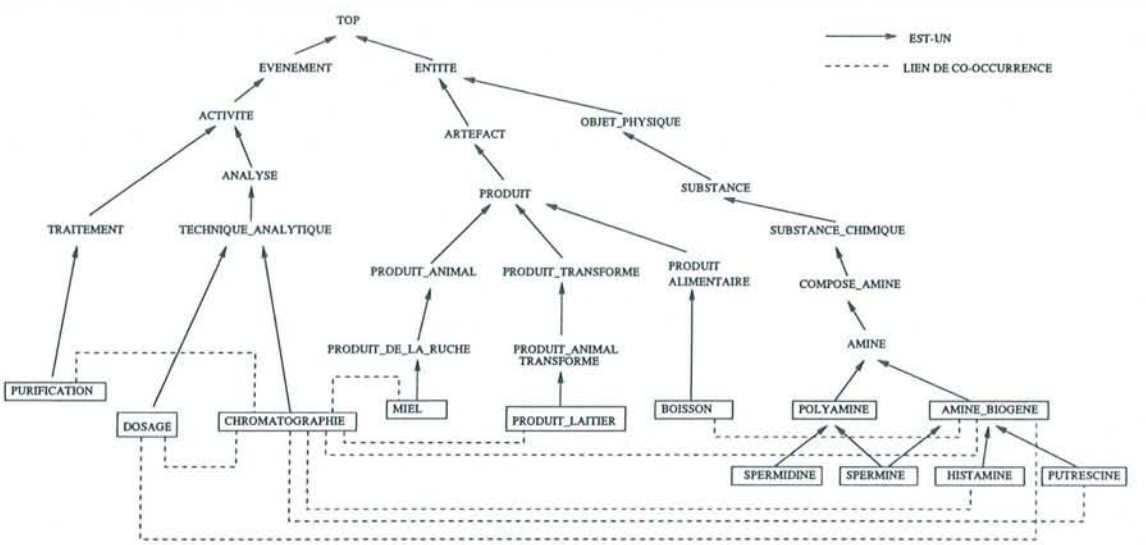

FIG. 6.6 – *Termes de la classe* CHROMATOGRAPHIE *replacés dans la hiérarchie, avec visualisation des liens de co-occurrences*

prédicatives mettant en jeu des termes de la classe qui permet de les identifier. Mais bien

plus que cela, les structures prédicatives vont permettre de généraliser les associations issues des classes de termes, et permettre d'accéder à une vue plus synthétique du contenu informationnel des textes.

6.3 Les structures prédicatives pour identifier les associations de type syntagmatique

Comme nous l'avons dit auparavant (section 5.4), nous nous limitons à l'identification de certaines structures prédicatives afin de limiter la complexité des traitements (cf. section 5.3.3). Tout d'abord, pour une classe de termes donnée, seules les structures prédicatives faisant intervenir au moins un terme de la classe sont prises en compte. Ensuite, nous laissons de coté les structures verbales, et ne nous intéressons qu'aux structures nominales. Ce choix est motivé par le fait que les textes scientifiques et techniques font largement appel à la nominalisation : un prédicat nominal est utilisé à la place d'un prédicat verbal, conduisant à un groupe nominal prédicatif. Ainsi, la phrase :

Les amines ont été dosés par chromatographie liquide haute pression.

peut être remplacée par la phrase suivante :

Un dosage des amines par chromatographie liquide haute pression a été effectué.

où l'information essentielle est exprimée sous forme nominale :

dosage des amines par chromatographie liquide haute pression.

La prédominance des phrases nominales sur les phrases verbales dans les textes scientifiques et techniques peut être plus ou moins marquée selon le type de textes, et le domaine considéré [Copeck 97]. Cette prédominance est vérifiée dans notre corpus.
Pour notre expérimentation, nous ne disposions pas de système d'extraction, et nous n'avions ni le temps ni les moyens nécessaires d'en concevoir un. Nous avons donc eu recours à une extraction non automatisée. Cependant, notre objectif étant de proposer un traitement « réaliste » des textes, nous avons volontairement restreint l'extraction aux groupes nominaux, afin de montrer notamment qu'elle est suffisante pour l'analyse de l'information.

Le tableau 6.3 illustre quatre groupes nominaux, extraits des résumés du corpus, ayant un comportement prédicatif. Ils sont composés d'un prédicat nominal et de ses arguments. La colonne de droite énumère les associations de la classe CHROMATOGRAPHIE qui correspondent à ces groupes nominaux. Les structures prédicatives correspondantes sont respectivement :

SP_1 : *dosage(objet : amine biogène, moyen : chromatographie liquide haute performance)*
SP_2 : *quantification(objet : produit laitier, moyen : chromatographie liquide haute pression)*
SP_3 : *détermination(objet : amine biogène, localisation : boisson)*
SP_4 : *purification(moyen : chromatographie sur gel)*
SP_5 : *dosage(objet : polyamines)*

#	Groupe nominal	type de lien	lien
GN_1	Dosage d'amines biogènes par chromatographie liquide haute performance	processus-objet, processus-processus	L_{10}, L_{12}, L_{15}
GN_2	Quantification de produits laitiers par chromatographie liquide haute pression	processus-objet	L_{13}
GN_3	Détermination d'amines biogènes dans les boissons	objet-objet	L_{11}
GN_4	Purification par chromatographie sur gel	processus-processus	L_{17}
GN_5	Dosage de polyamines	processus-objet	

TAB. 6.3 – *Exemples de groupes nominaux du corpus illustrant les liens de co-occurrence*

Les structures prédicatives permettent de caractériser les liens de co-occurrence de la classe de termes. Ainsi, le groupe nominal GN_1 et sa structure prédicative correspondante SP_1 apportent des précisions sur trois liens : L_{10}, L_{12} et L_{15}. Le lien L_{10} entre *amine biogène* et *chromatographie* s'explique par l'action de dosage d'une substance chimique (l'*amine*) par une certaine technique analytique (la *chromatographie*). La structure prédicative permet de re-situer l'association entre les deux termes dans son contexte.

L'intérêt principal du recours aux structures prédicatives est de pouvoir généraliser les associations d'une classe. En effet, en prenant en compte la hiérarchie des termes, il est possible de regrouper des termes de sens proches[63] pour en déduire un terme plus général. Par exemple, considérons le lien de co-occurrence (L_{13}), qui relie *produit laitier* et *chromatographie*. Ceci est reflété par la structure prédicative SP_2. Par ailleurs, une autre structure prédicative du corpus montre un lien entre *miel* et *chromatographie*. Il est alors légitime de considérer qu'il existe un lien plus général entre les deux termes *produit* et *chromatographie* comme l'illustre la figure 6.7.
De la même façon, les groupes nominaux GN_1 et GN_5, mettent en jeu les deux termes *polyamine* et *amine biogène* avec le terme *dosage* : nous pouvons induire un lien plus général entre les termes *amine* et *dosage*, *amine* correspondant à la généralisation de *polyamine* et *amine biogène* (figure 6.8).
Ceci revient à effectuer une généralisation sur les structures prédicatives. Notre processus de généralisation, présenté au chapitre 4, peut ainsi être appliqué sur les structures prédicatives extraites et permettre de synthétiser le contenu informationnel correspondant. Les structures prédicatives extraites et les généralisations obtenues forment un moyen pertinent de collecter l'information des textes, à mi-chemin entre les classes de termes et les textes.

Les règles d'extraction des structures prédicatives à partir du corpus ont été les suivantes :

- la structure prédicative doit être sous forme nominale,
- les arguments, ou une partie des arguments, doivent avoir été reconnus comme termes par le système FASTER,
- les rôles thématiques utilisés sont : agent, objet, but, moyen, localisation.

63. C'est-à-dire dont les positions dans la hiérarchies sont proches.

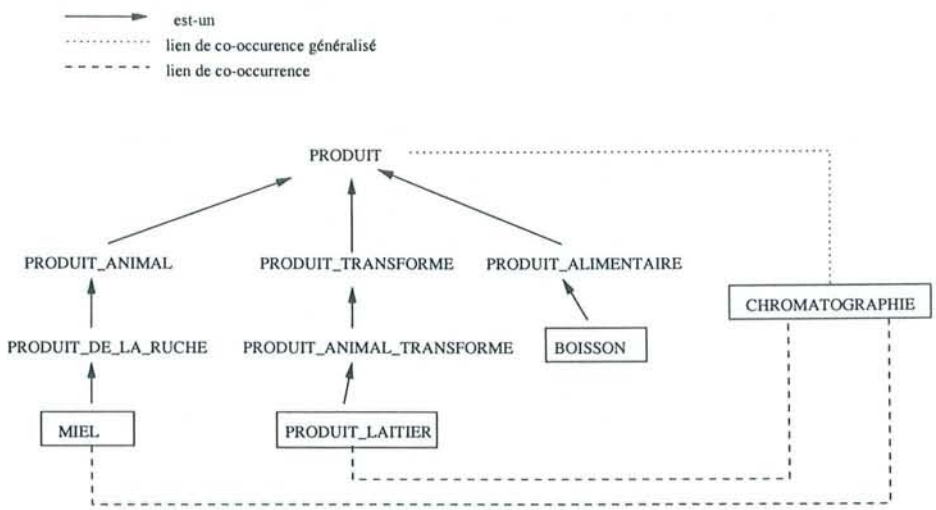

FIG. 6.7 – *Généralisation des liens entre* `chromatographie, produit laitier, miel`

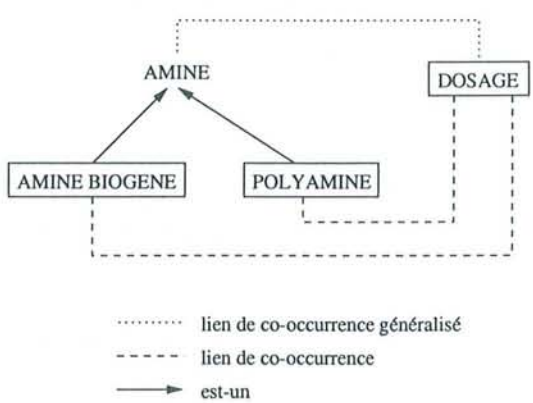

FIG. 6.8 – *Généralisation des liens entre* `dosage, amine biogène, polyamine`

Les prédicats nominaux extraits ont été introduits dans la hiérarchie lorsqu'ils n'en faisaient pas partie.

Nous présentons dans la section suivante notre prototype qui met en oeuvre la généralisation des structures prédicatives.

6.4 Un prototype pour la prise en compte des structures prédicatives

Nous avons mis en oeuvre un prototype permettant d'effectuer la généralisation des structures prédicatives et de visualiser le résultat. Nous situons ce prototype comme un élément supplémentaire dans la chaîne de traitement de l'analyse de l'information, comme outil complémentaire aux cartes thématiques fournies par SDOC. Il est donc destiné à un utilisateur réalisant l'analyse d'un ensemble de textes.

Le prototype est composé de deux modules :

- un noyau permet de représenter les structures prédicatives au sein d'une hiérarchie et de calculer des généralisations,
- une interface utilisateur permet d'exploiter et de visualiser les différentes informations (concepts, structures prédicatives).

Le noyau utilise la logique de descriptions CLASSIC, qui fournit les fonctionnalités de base pour la représentation des structures prédicatives et la gestion de la hiérarchie. Nous avons conçu un ensemble de fonctions Lisp mettant en oeuvre les algorithmes de généralisation que nous avons présentés au chapitre 4.

L'interface utilisateur permet :

1 de visualiser la hiérarchie selon plusieurs modalités : (1) avec ou sans les structures prédicatives, à des niveaux de profondeur variables ; (2) en se focalisant sur les termes d'une classe particulière,

2 de visualiser la liste des prédicats utilisés, des structures prédicatives initiales, des structures prédicatives calculées (généralisations),

3 de visualiser la description complète d'une structure prédicative ou d'un concept de la hiérarchie,

4 de parcourir les structures prédicatives à l'aide d'une fonction hypertexte, selon l'ordre hiérarchique.

La visualisation de la hiérarchie et ses différents modes est illustrée par les figure 6.9, 6.10 et 6.11. La première vue (figure 6.9) permet de se familiariser avec la hiérarchie des concepts. La deuxième (figure 6.10) permet une vue globale, sans détail, des différentes généralisations[64]. La troisième vue (figure 6.11) permet de situer tous les termes d'une classe sur un seul graphe, à partir du nom de la classe de terme. La figure 6.12 montre la liste des généralisations calculées sur la classe chromatographie, dont on peut consulter la description comme cela est illustré par la figure 6.13.

La visualisation sous forme hypertexte est destinée à obtenir une vision hiérarchisée du contenu informationnel. Le point d'entrée est constitué par les structures prédicatives les

64. Le terme *dosage* apparaît sous sa forme préférentielle donnée par le thésaurus AGROVOC, *analyse quantitative*. En effet, dans AGROVOC, *dosage* est un synonyme du terme *analyse quantitative*.

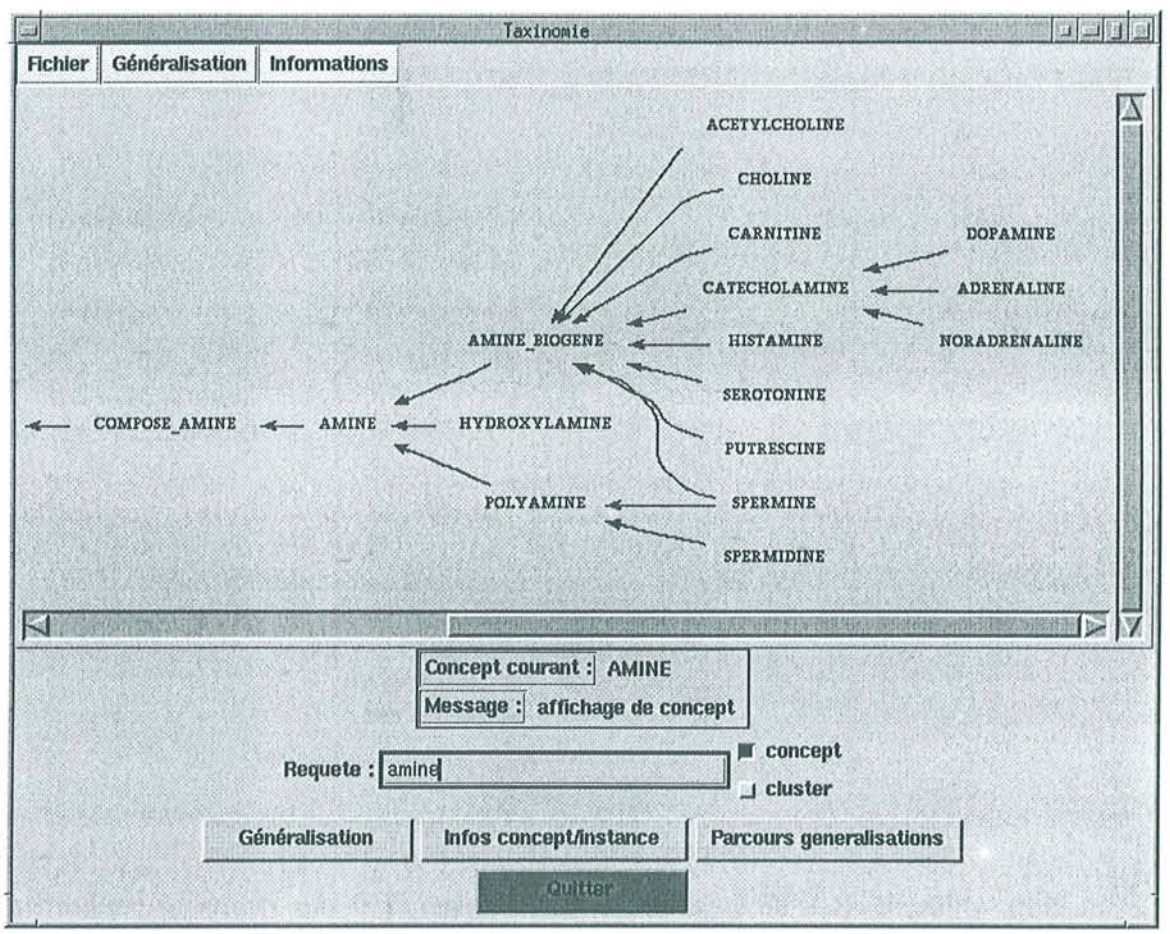

FIG. 6.9 – *Visualisation de la hiérarchie des concepts, sans structures prédicatives*

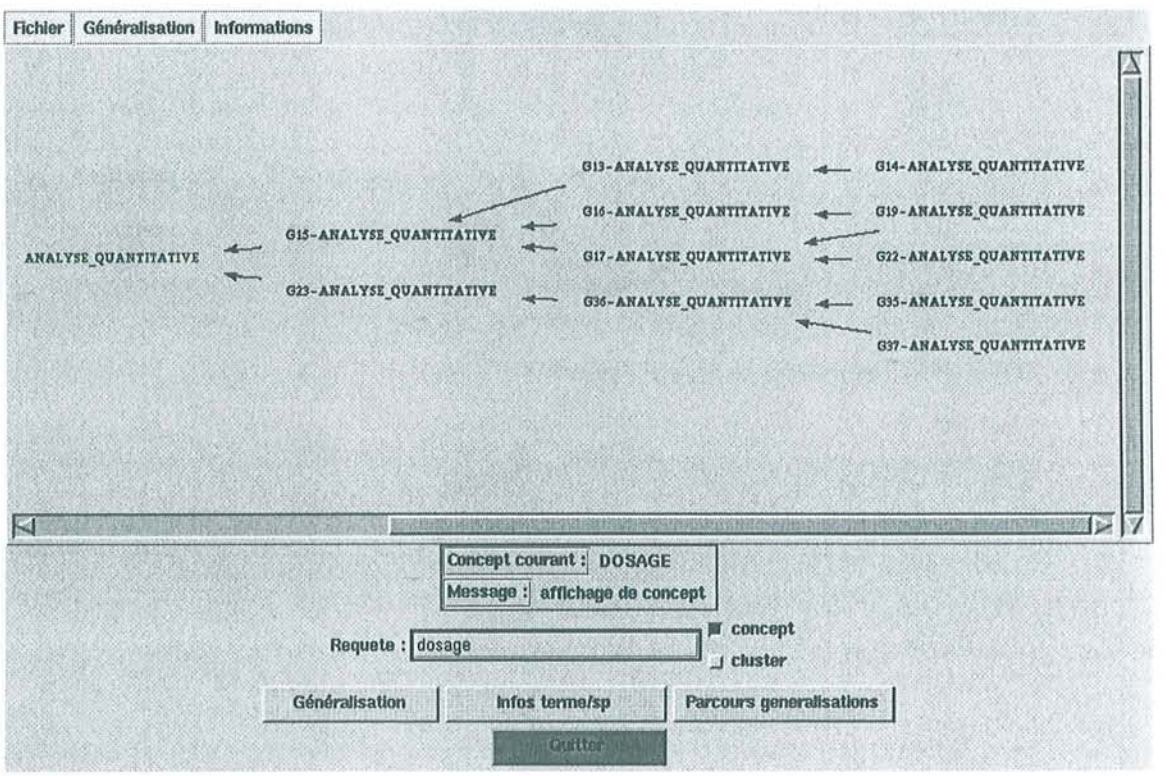

FIG. 6.10 – *Visualisation de la hiérarchie des concepts, avec des structures prédicatives (généralisations préfixées par la lettre « G »)*

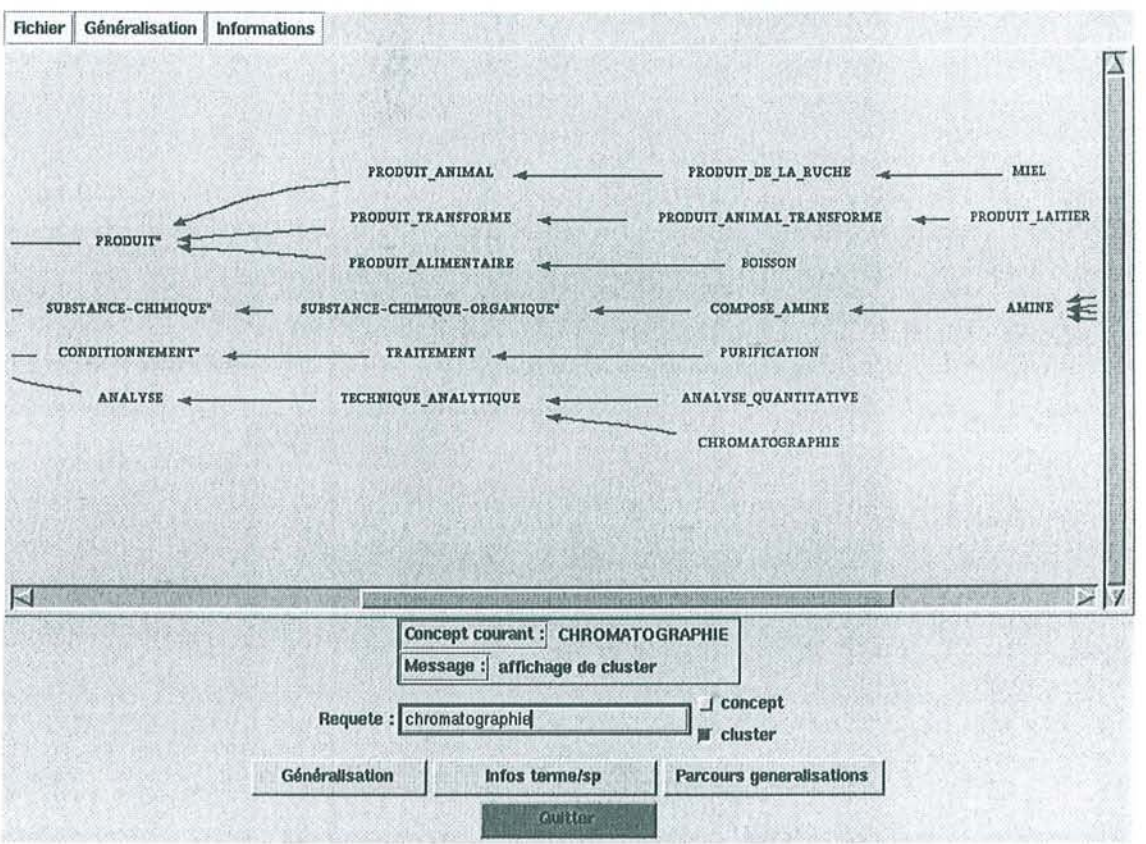

FIG. 6.11 – *Visualisation des termes de la classe* CHROMATOGRAPHIE *projetés sur la hiérarchie*

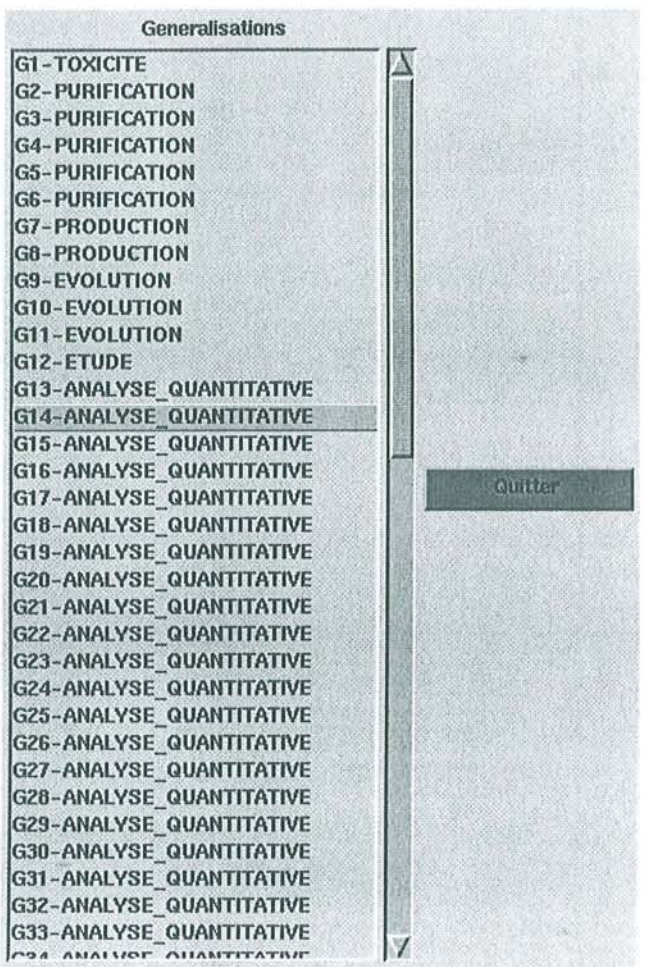

FIG. 6.12 – *Visualisation de la liste des généralisations calculées*

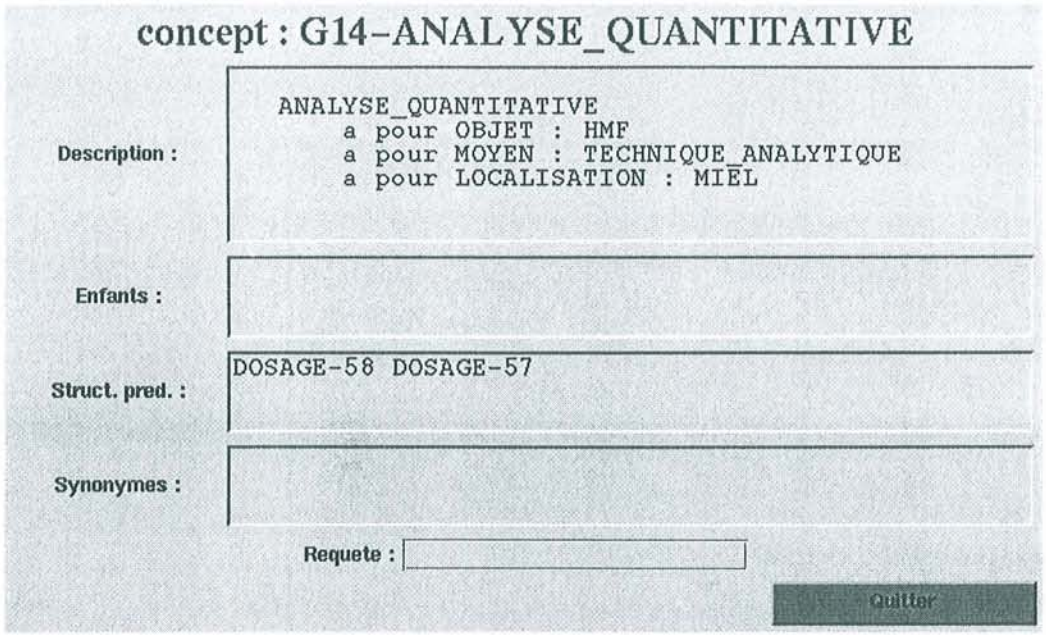

FIG. 6.13 – *Visualisation de la description d'une structure prédicative, obtenue par double-clic sur un élément de la figure 6.12*

plus générales calculées pour une classe donnée. Ces structures sont affichées sous forme textuelle accompagnées de leur description. En cliquant sur la zone d'une structure prédicative A, on accède à ses fils, qui sont à leur tour affichés. Il est ainsi possible de parcourir les structures prédicatives des plus générales aux plus spécifiques, pour découvrir les différentes informations extraites des textes. Pour chaque structure prédicative affichée, nous donnons des informations numériques permettant de connaître l'importance quantitative d'une information dans les textes. Ces informations numériques sont de deux types :

- pour une structure prédicative A, nous donnons le pourcentage de structures prédicatives couvertes par rapport à l'ensemble des structures prédicatives collectées,
- pour une structure prédicative A, de parent B, nous donnons le pourcentage de structures prédicatives couvertes par A par rapport aux structures prédicatives couvertes par B.

Les figures 6.14 et 6.15 montrent deux écrans de parcours hypertexte des structures prédicatives. La première figure (6.14) concerne la généralisation `G16-ANALYSE_QUANTITATIVE` dont la description (absente de la figure) est :

```
G16-ANALYSE_QUANTITATIVE ≐ (and ANALYSE_QUANTITATIVE
                                (all moyen CHROMATOGRAPHIE))
```

A côté du nom, `G16-ANALYSE_QUANTITATIVE`, figure en absolu et en pourcentage le nombre de structures prédicatives subsumées. Ainsi, `G16-ANALYSE_QUANTITATIVE` couvre 9 structures prédicatives sur 124 extraites, ce qui représente 7,3% des structures prédicatives. Les informations données sont les structures prédicatives qui sont fils du concept `G16-ANALYSE_QUANTITATIVE`.
On observe ainsi qu'il y a une généralisation, `G19-ANALYSE_QUANTITATIVE`, et trois structures prédicatives initiales, `DOSAGE-30`, `DOSAGE-28` et `DOSAGE-104`, qui sont des subsumés

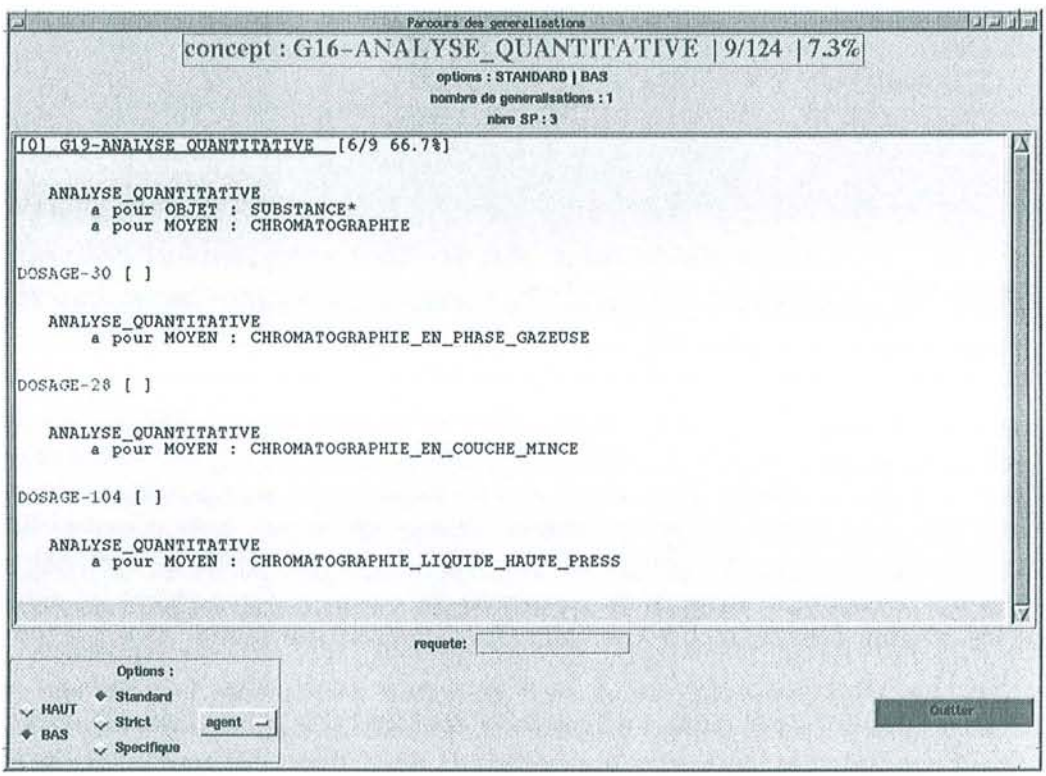

FIG. 6.14 – *Écran de parcours hypertexte des structures prédicatives, faisant apparaître les fils de la généralisation* `G16-ANALYSE_QUANTITATIVE`

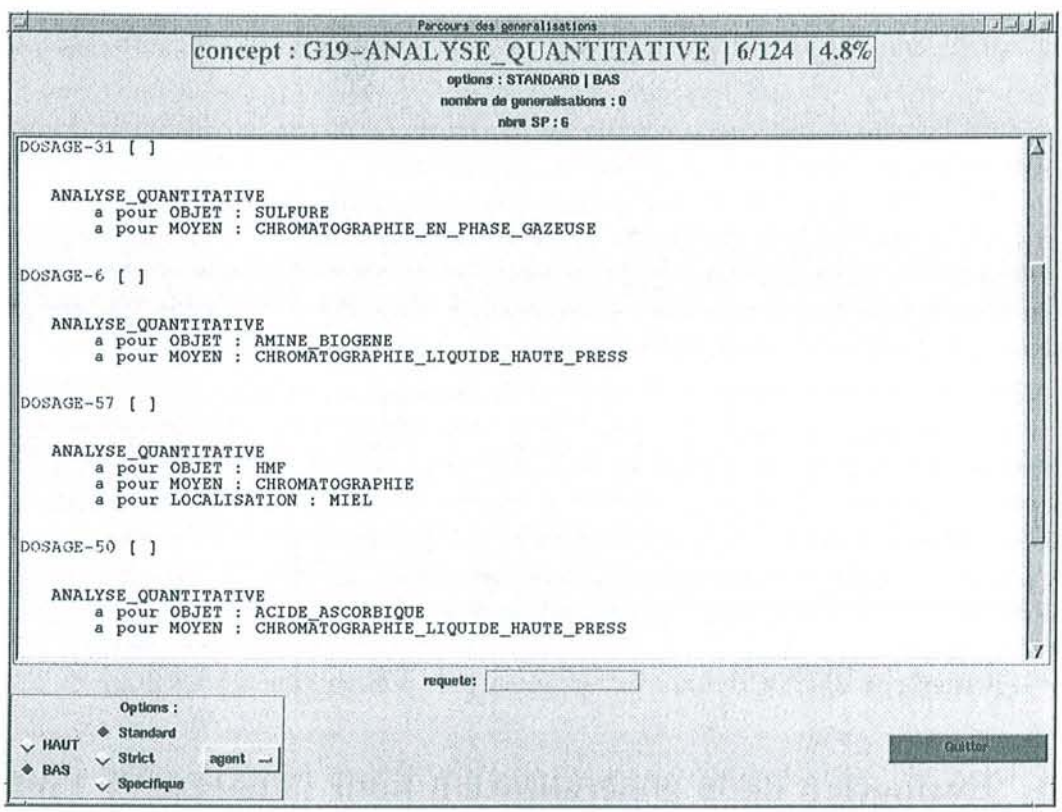

FIG. 6.15 – *Écran de parcours hypertexte des structures prédicatives, faisant apparaître les fils de la généralisation* `G19-ANALYSE_QUANTITATIVE`

immédiats de G16-ANALYSE_QUANTITATIVE. La description de chaque structure prédicative (généralisation ou structure initiale) est affichée. Si une structure possède des descendants, comme c'est le cas de G19-ANALYSE_QUANTITATIVE sur la figure, il est possible d'accéder par simple clic sur sa zone à ses descendants directs. De plus, le nombre de structures prédicatives auquel elle permet d'accéder est affiché, en absolu et en pourcentage par rapport aux structures couvertes par le concept courant (ici G16-ANALYSE_QUANTITATIVE). Dans le cas de G19-ANALYSE_QUANTITATIVE, 6 des 9 structures prédicatives couvertes par G16-ANALYSE_QUANTITATIVE sont accessibles, soit un peu plus de 66%.

La deuxième figure (6.15) concerne la généralisation G19-ANALYSE_QUANTITATIVE, obtenue à partir de l'écran présentant G19-ANALYSE_QUANTITATIVE en cliquant sur la zone textuelle correspondante. G19-ANALYSE_QUANTITATIVE subsume directement 6 structures prédicatives, dont 4 sont visualisées sur l'écran. Elle ne subsume aucune autre généralisation.

L'outil hypertexte est conçu comme un moyen de découvrir le contenu informationnel des textes de manière descendante, en commençant par les structures prédicatives les plus générales, couvrant de nombreuses structures prédicatives initiales, pour aboutir à des informations très spécifiques. L'utilisateur peut découvrir ainsi progressivement les événements relevés dans les textes, sans être perdu dans une masse d'informations. Il est également possible de remonter la hiérarchie et d'accéder directement aux structures prédicatives en donnant une requête sous la forme d'un simple prédicat.

Notre interface est limitée au niveau des fonctionnalités, car nous avons consacré peu de temps à sa mise en oeuvre. En particulier, nous n'avons pas intégrées celles qui sont offertes par les outils de visualisations de SDOC (dont le nom est HENOCH [Grivel 97]). Il n'est donc pas possible à l'aide de notre prototype d'accéder à la liste des termes et des associations pour une classe donnée, ainsi qu'à la carte thématique, mais surtout d'accéder aux documents à partir d'un terme ou d'une association. La mise en oeuvre de ces fonctionnalités dans notre prototype ne pose cependant aucune difficulté.
L'intérêt de notre prototype réside dans la prise en compte des structures prédicatives, ce que ne permet pas HENOCH, qui n'est qu'un outil de présentation des résultats de SDOC.

6.5 Evaluation de la généralisation pour l'analyse de l'information

Pour évaluer l'apport de la généralisation au processus d'analyse de l'information, nous avons réalisé une expérimentation avec un expert documentaliste[65]. L'expérimentation a porté sur l'analyse de deux classes de termes : l'une avec l'environnement fourni par la plateforme ILIAD, et l'autre avec notre prototype de généralisation de structures prédicatives. L'objectif est de cerner l'apport du processus de généralisation pour l'analyse finale réalisée par l'expert.
Dans un premier temps, nous montrons que la qualité des généralisations obtenues est difficile à établir, car elle est assez subjective. Nous essayons d'identifier les paramètres qui peuvent faire varier cette qualité. Nous exposons ensuite l'expérimentation avec l'expert documentaliste, et détaillons les résultats, puis nous concluons sur cette évaluation.

65. Nous appelons ainsi un documentaliste qui possède une bonne connaissance (une expertise) d'un domaine de spécialité.

6.5.1 La qualité d'une généralisation, une notion très relative

La hiérarchie obtenue peut être jugée selon la qualité des généralisations qu'elle contient. Cette notion est difficile à définir dans la mesure où il n'existe pas *a priori* de critères objectifs. Nous essayons de définir des critères généraux puis d'identifier les facteurs qui vont intervenir dans l'obtention de généralisations satisfaisantes.

Une généralisation doit être un bon compromis entre généralité et précision.
Elle doit être suffisamment générale : elle est peu utile si elle ne permet pas de synthétiser plusieurs structures prédicatives. Par exemple, si les généralisations ne correspondent qu'à la synthèse de structures prédicatives deux à deux, la hiérarchie résultante possède un grand nombre de structures intermédiaires qui rendent difficile l'interprétation du contenu informatif. La généralité des structures permet d'assurer une bonne couverture des informations. Par exemple, une structure prédicative qui couvre 10% ou 20% de l'ensemble des structures prédicatives est très utile pour cerner le contenu informatif global.

Une généralisation doit être suffisamment précise : si les termes utilisés sont trop abstraits, trop généraux, alors l'information est quasiment nulle. Elle doit apporter une information. Les concepts du haut de la hiérarchie, très abstraits, tel que « objet physique » ou « événement » n'apportent finalement aucune information pertinente.

Deux facteurs essentiels interviennent dans la qualité d'une généralisation, indépendants de l'algorithme de calcul utilisé :

- la qualité de la hiérarchie de concepts utilisée,
- la diversité des informations contenues dans les structures prédicatives que l'on cherche à généraliser.

La hiérarchie de concepts initiale, utilisée pour catégoriser les différentes connaissances sur le domaine considéré, influe beaucoup sur les généralisations obtenues. Si elle est trop approximative, cela se répercute dans les généralisations. Idéalement, il faut donc disposer d'une hiérarchie de concepts adaptée à l'utilisateur final, à sa terminologie et à l'image mentale qu'il se fait du domaine. Toute différence entre la hiérarchie telle qu'elle est conçue et l'idée que s'en fait l'utilisateur introduit un facteur d'incompréhension qui réduit l'intérêt de disposer d'une vue synthétique du contenu informationnel des textes. Par exemple, si la hiérarchie positionne le concept A comme un fils du concept B, et que l'utilisateur perçoit A et B comme des concepts de même niveau, il ne comprendra pas la généralisation de A et B en A, et s'attendra à trouver un concept C subsumant à la fois A et B.

La diversité des informations contenues dans les structures prédicatives joue également un rôle prépondérant. C'est un facteur externe, qui dépend finalement de l'homogénéité du corpus de textes traité. Si les structures prédicatives sont très hétérogènes, la généralisation conduit à un ensemble de structures prédicatives très générales, constituées de termes trop généraux pour être pertinents. Calculer des généralisations consiste en quelque sorte à appauvrir les descriptions individuelles afin d'en déduire des descriptions plus globales mais moins précises. Si cet appauvrissement est trop grand, les descriptions perdent leur intérêt, et deviennent inutiles. Une autre conséquence de l'hétérogénéité des informations est une augmentation du nombre de niveaux hiérarchiques, qui contribue à rendre moins lisible la hiérarchie.

En ce qui concerne notre expérimentation, le corpus, bien que couvrant de nombreux thèmes de l'agriculture, s'avère assez homogène. 50 classes ont été calculées par SDOC, et

nous avons choisi une classe bien positionnée (centralité et densité). Il n'est évidemment pas possible d'obtenir le même résultat avec les 50 classes. En ce qui concerne la hiérarchie initiale de concepts, elle n'est pas d'une excellente qualité : d'une part, elle provient de la réutilisation d'un thésaurus destiné à la consultation humaine, donc peu adapté à priori à une classification rigoureuse ; d'autre part, nous n'avons pas pu[66] nous concerter avec l'expert documentaliste afin d'adapter la hiérarchie à son expertise, et nous ne sommes pas spécialiste du domaine.

6.5.2 Une expérimentation avec un documentaliste expert

L'expérimentation consiste à comparer deux méthodes pour analyser le contenu des documents indexés par une classe de termes issue de SDOC.
La première méthode utilise l'outil de visualisation classique de SDOC, HENOCH, qui permet de consulter la carte thématique, les associations entre termes, ainsi que les titres et les documents du corpus. Cependant elle ne permet pas l'utilisation de structures prédicatives.
La deuxième méthode utilise les outils de visualisation de notre interface, et permet d'utiliser les structures prédicatives et leur généralisations.

Nous avons utilisé des classes de termes différentes pour chaque méthode, ayant des coefficients de centralité et de densité voisins : la classe CERNE et la classe CHROMATOGRAPHIE.

Analyse de la classe CERNE sans structures prédicatives

La méthode utilisée par l'INIST pour analyser les classes de termes exploite l'outil de visualisation associé à SDOC, HENOCH. Le point d'entrée est constitué par la liste des classes de termes calculées ou alternativement par la carte thématique déjà illustrée lors de la présentation de SDOC (cf. section 5.3.2, figure 5.3).

La navigation au sein de HENOCH se fait par hypertexte, par simple clic sur les différents éléments textuels. Dans notre cas, nous nous intéressons à la classe de termes CERNE, dont la description est représentée par les deux copies d'écran des figures 6.16 et 6.17. Le documentaliste expert est capable de repérer globalement les différents thèmes de la classe à partir de la liste des termes de celle-ci. Dans le cas de CERNE, il analyse ainsi la classe comme portant sur la foresterie et l'état sanitaire des arbres, et notamment le moyen de déterminer la vigueur des arbres, par l'intermédiaire des cernes qui renseignent sur le passé de l'arbre et sur ce qui a influencé son développement.

Après une première caractérisation basée sur les termes, le documentaliste vérifie si d'autres classes correspondent aux thèmes de la forêt, par l'intermédiaire des associations externes (bas de la figure 6.17) ou de la carte thématique. Il repère par exemple deux classes, ARBRE et FORESTIER, qui ont certains points communs avec CERNE, mais qui sont focalisés sur d'autres thèmes comme les techniques de sylviculture.

Il consulte ensuite, une par une, les différentes associations de la classe, données par HENOCH (figure 6.17). D'après le documentaliste, les associations permettent de recons-

66. Par manque de temps, la disponibilité de l'expert étant réduite.

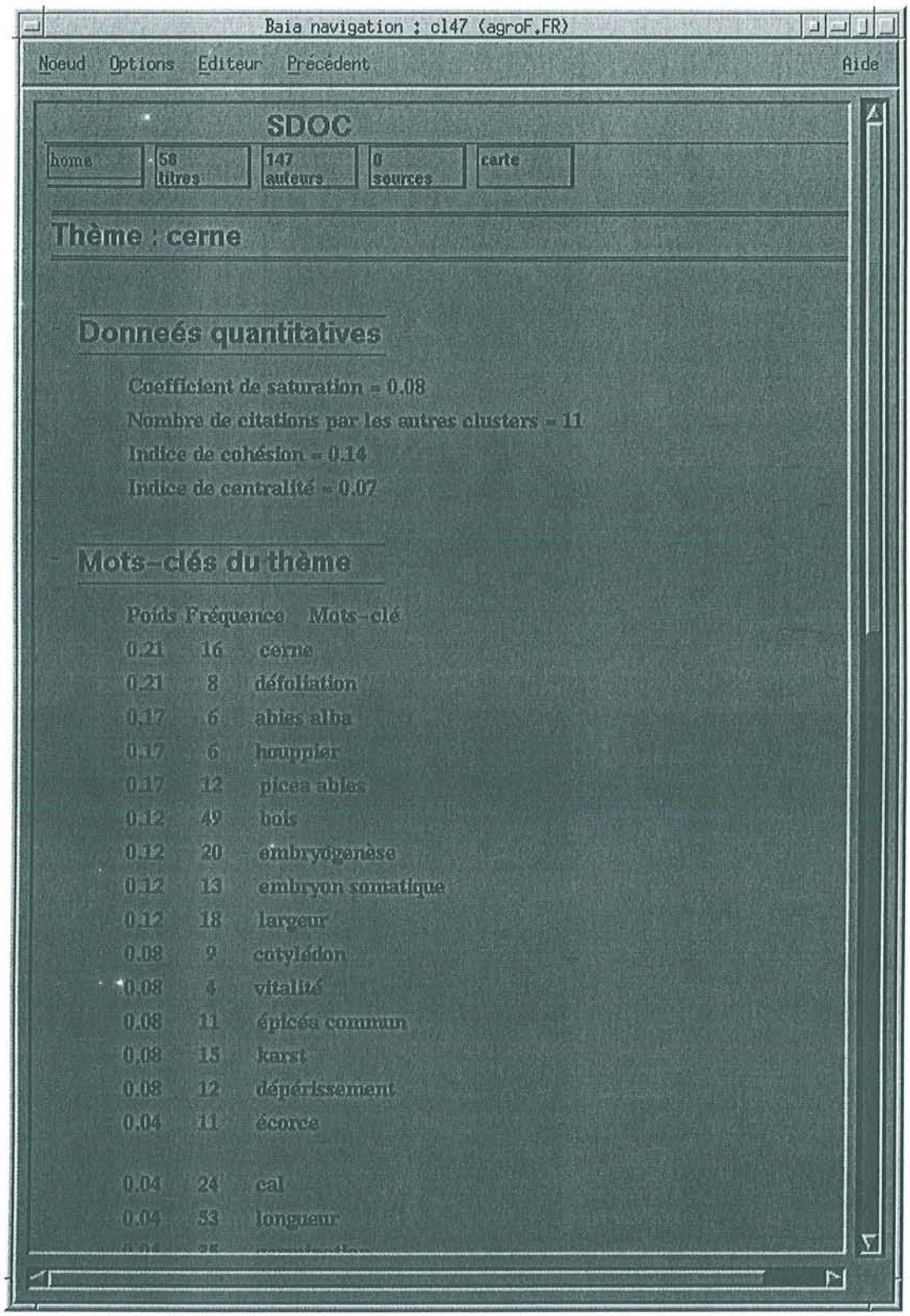

FIG. 6.16 – *Visualisation d'une classe de termes par l'interface de SDOC : termes de la classe CERNE*

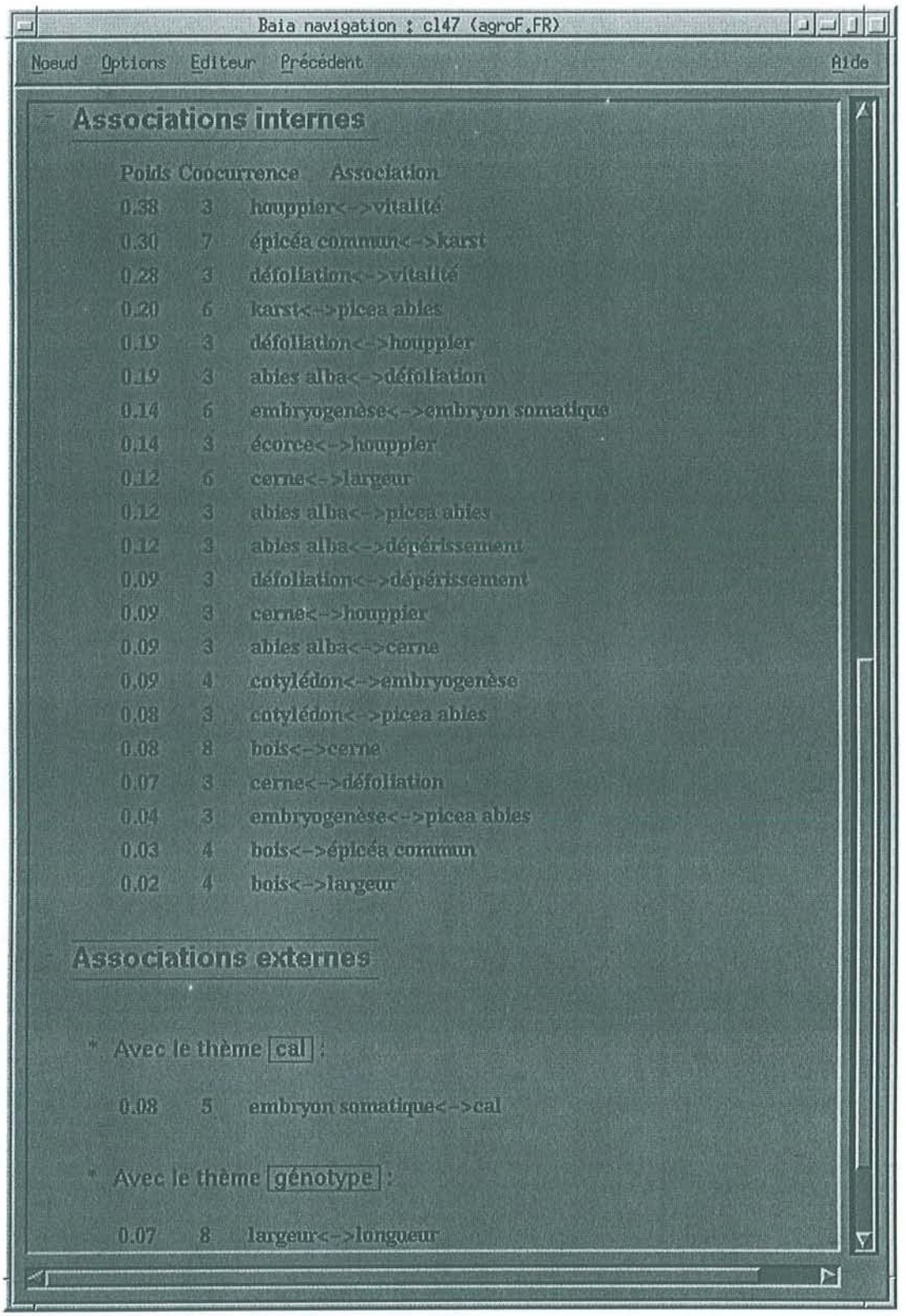

FIG. 6.17 – *Visualisation d'une classe de termes par l'interface de SDOC : associations internes et externes de la classe CERNE*

tituer les chaînes de mots qui se trouvent dans les textes. L'analyse des associations fait apparaître une grande diversité quant aux informations sous-jacentes. Il existe ainsi :

- des associations que le documentaliste considère comme tout à fait normales, c'est-à-dire qui sont attendues et ne constituent pas un élément original. Elles viennent confirmer une connaissance bien connue de l'expert, comme par exemple l'association entre les termes `houppier`[67] et `vitalité` ou l'association entre les terme `défoliation` et `vitalité`,
- des associations qui résultent d'un artefact, et qui n'ont aucune signification réelle. Par exemple, l'association entre `picea abies`[68] et `karst` provient ici d'un nom propre `Picea abies Karst` apparaissant dans plusieurs documents, et non d'une association réelle avec le terme `karst`,
- des associations qui sont en dehors du thème représenté par la classe, et peu en rapport avec le reste. Par exemple, l'association entre `embryogénèse` et `embryon somatique`, qui concerne une technique d'embryogénèse et non la vitalité des arbres,
- des associations polysémiques, qui recouvrent en fait plusieurs aspects qui ne sont pas différenciés, comme c'est le cas de l'association entre `germination` et `embryogénèse`, qui n'est pas forcément liée aux graines.

Lors de l'analyse des termes ou des associations, le documentaliste est souvent amené à consulter, pour obtenir plus de précisions sur le sens de ces derniers :

- soit les titres des documents, obtenus en cliquant sur le terme ou l'association concernée,
- soit un ou plusieurs documents, obtenus à partir des titres ou directement à partir des termes ou associations.

Il y a donc deux niveaux d'informations disponibles pour vérifier le sens d'un terme ou d'une association. Les titres permettent de situer l'association dans un contexte limité à une phrase. L'avantage des titres est qu'il est possible de les afficher dans une fenêtre unique : l'accès est donc très rapide. La figure 6.18 montre la visualisation des titres de la classe CERNE. En pratique, il est possible d'afficher les seuls titres des documents correspondants à un terme ou à une association, ce qui réduit leur nombre. Parfois le titre est insuffisant, notamment lorsque l'association n'y apparaît pas ou lorsque le contexte n'est pas suffisamment précis. Le documentaliste doit alors accéder aux documents complets. La visualisation est moins rapide dans la mesure où chaque document est représenté dans une fenêtre séparée. La figure 6.19 illustre la représentation d'un document entier.

Toutes ces informations permettent au documentaliste d'analyser le contenu informationnel des documents indexés par la classe de terme et d'en déduire un thème central qui sera retenu pour la rédaction du rapport de tendance.
Le documentaliste est capable de verbaliser de nombreuses connaissances à partir des termes et des associations d'une classe. Toutefois l'analyse nécessite une part importante de consultation, à la fois des titres et des documents, pour vérifier certaines hypothèses ou préciser des éléments d'informations. Nous montrons dans la section suivante que l'utilisation de structures prédicatives permet une analyse plus fine sans nécessiter un retour aux documents.

67. Le houppier est la couronne de l'arbre, et sa qualité est directement en rapport avec la vitalité de l'arbre.
68. Nom d'une espèce d'arbres, l'épicéa commun.

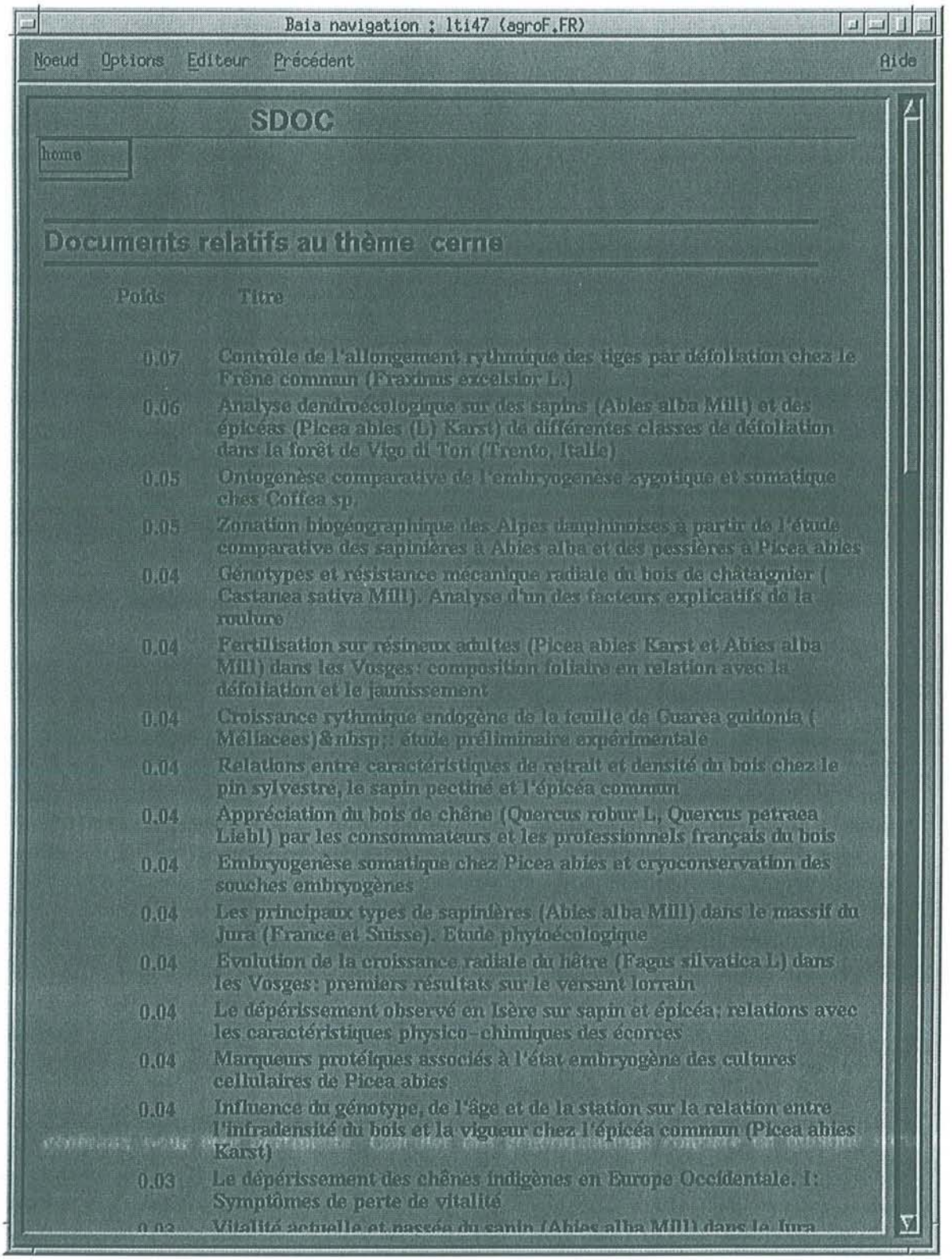

FIG. 6.18 – *Visualisation des titres des documents relatifs à la classe CERNE*

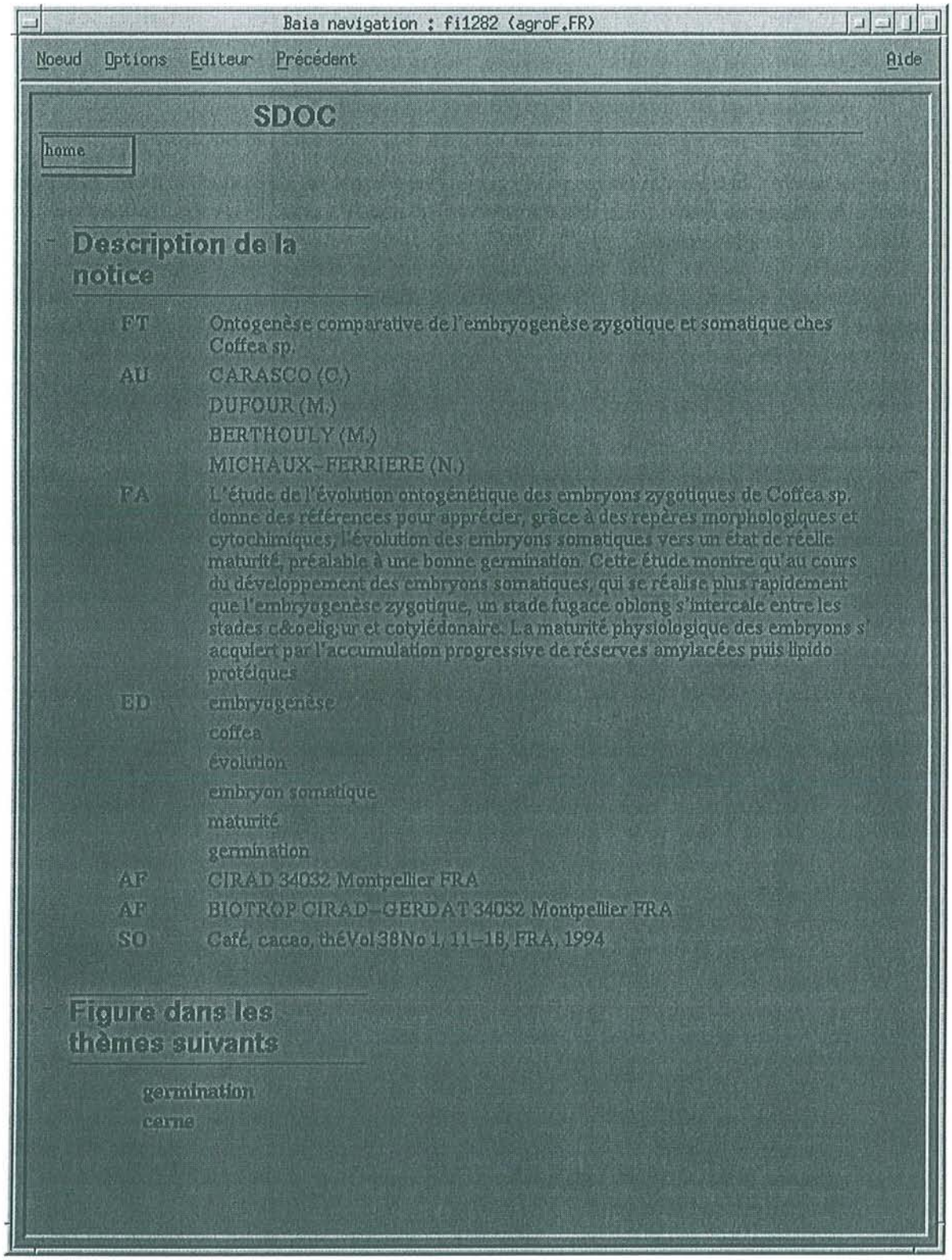

FIG. 6.19 – *Visualisation de la description et du contenu d'un document relatif à la classe CERNE*

Analyse de la classe CHROMATOGRAPHIE avec des structures prédicatives

Notre méthode est destinée à améliorer l'accès au contenu informationnel :

- en évitant ou en diminuant le retour aux documents,
- en offrant une vue synthétique du contenu d'un ensemble de documents.

L'hypothèse est que les structures prédicatives remplissent un rôle informatif suffisant pour éviter le retour au texte plein des documents, et que le calcul de généralisations permet d'éviter un accés séquentiel.
Dans cette perspective, nous avons tenté de dégagé les améliorations apportées par cette nouvelle méthode et les problèmes spécifiques qu'elles posent. Nous ne mettons donc pas en avant l'analyse comme nous l'avons fait pour la première méthode avec HENOCH, mais les réflexions de l'expert documentaliste sur notre proposition. Nous insistons sur le fait que la présentation que nous avons faite a été limitée du fait que les fonctionnalités de l'interface HENOCH n'ont pas été intégrées à notre prototype, ce qui restreint les possibilités de visualisation.

Les structures prédicatives permettent une analyse plus fine et non séquentielle

De manière générale, le documentaliste a perçu les structures prédicatives comme des informations pertinentes. Elles lui permettent une meilleure caractérisation des associations sans nécessiter l'affichage complet des documents. Le recours aux titres (non accessible avec notre prototype) reste une alternative pertinente à la visualisation des structures prédicatives car il n'est pas aussi pénalisant que l'accès aux textes, qui nécessite de rechercher le passage correspondant à l'information recherchée.

Le documentaliste peut par exemple, grâce au parcours des structures prédicatives possédant le prédicat *dosage*, observer les différentes techniques d'analyse quantitative utilisées. La figure 6.20 que nous avons déjà présentée auparavant (figure 6.14) et que nous reproduisons ici montre trois structures prédicatives (`DOSAGE-30`, `DOSAGE-28`, `DOSAGE-104`) faisant apparaître trois techniques différentes de chromatographie pour le dosage. Des structures prédicatives plus précises, caractérisées par un rôle *objet* et un rôle *moyen*, sont synthétisées par la généralisation `G19-ANALYSE-QUANTITATIVE`. Un simple clic permet au documentaliste d'accéder aux descriptions correspondantes, illustrées par la figure 6.21 (reproduction de la figure 6.15), et détaillant les différentes substances utilisées. Ainsi le dosage s'applique-t-il au sulfure, aux amine biogènes, à l'HMF, à l'acide ascorbique, ... Ces informations sont quantifiées, ce qui permet au documentaliste d'apprécier leur importance au sein du corpus. Les informations examinées ici ne peuvent être collectées avec HENOCH qu'en examinant de nombreux documents.

Le parcours des structures prédicatives permet ainsi :

- de mettre en évidence de nombreuses unités informations disséminées au sein des documents. Lorsque le documentaliste consulte un document, il cherche des informations précises qu'il doit localiser. La donnée des structures prédicatives réduit fortement cette recherche,
- de structurer ces informations selon leur proximité sémantique, exprimée par des prédicats communs et des arguments proches dans la hiérarchie. Avec SDOC et l'interface HENOCH, cette synthèse des informations doit être réalisée par le documentaliste et nécessite de couvrir la majorité des documents, de manière séquentielle, pour ne pas omettre d'informations.

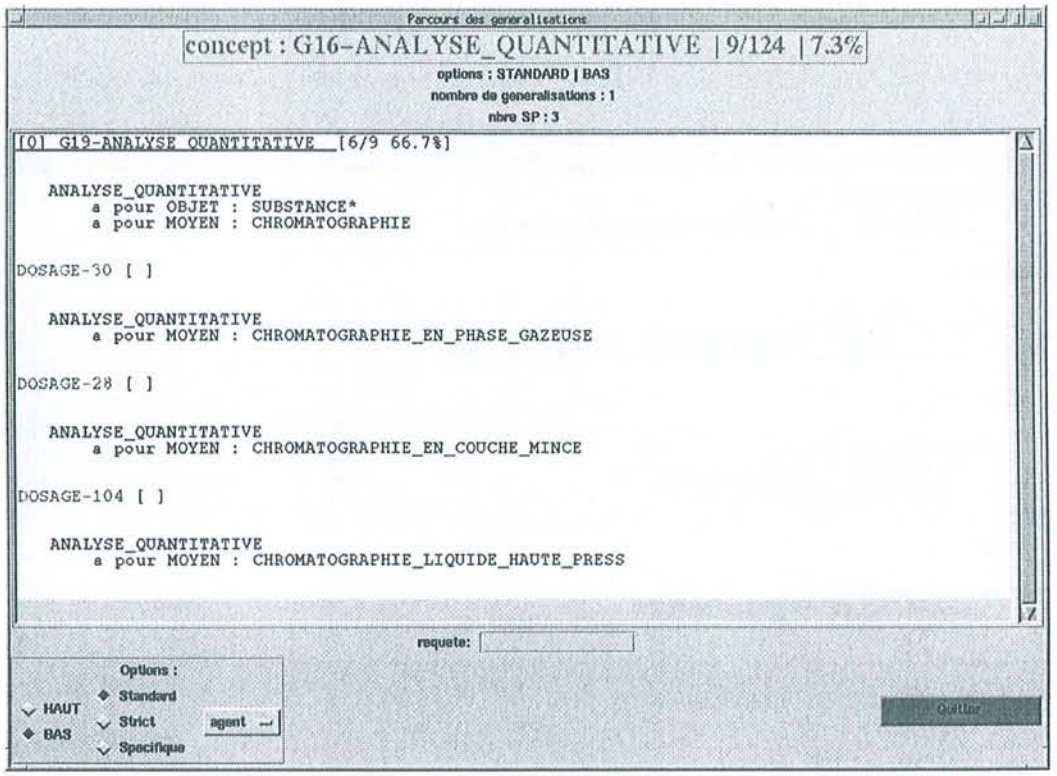

FIG. 6.20 – *Écran de parcours hypertexte des structures prédicatives, faisant apparaître les fils de la généralisation* ***G16-ANALYSE_QUANTITATIVE***

Les structures prédicatives permettent une synthèse pertinente de la classe de termes

Outre la navigation hiérarchique parmi les structures prédicatives initiales et calculées, il est possible d'exploiter plus finement les quelques structures prédicatives qui permettent une synthèse pertinente de la classe. Pour la classe CHROMATOGRAPHIE, les trois généralisations les plus intéressantes sont les suivantes :

```
G19-DOSAGE ≐ (and DOSAGE
                  (all objet SUBSTANCE_CHIMIQUE)
                  (all moyen ACTIVITE)
                  (all localisation PRODUIT))

G22-DOSAGE ≐ (and DOSAGE
                  (all objet SUBSTANCE_CHIMIQUE)
                  (all moyen CHROMATOGRAPHIE))

G36-DOSAGE ≐ (and DOSAGE
                  (all objet AMINE))
```

Nous avons extraits ces généralisations manuellement pour les présenter au documentaliste. Elles permettent de rendre compte globalement de toutes les associations de la classe. La généralisation G19-DOSAGE met à jour quatre concepts importants de la classe : le *dosage*, qui est le processus principal ; les *produits* sur lesquels le processus est appliqué ;

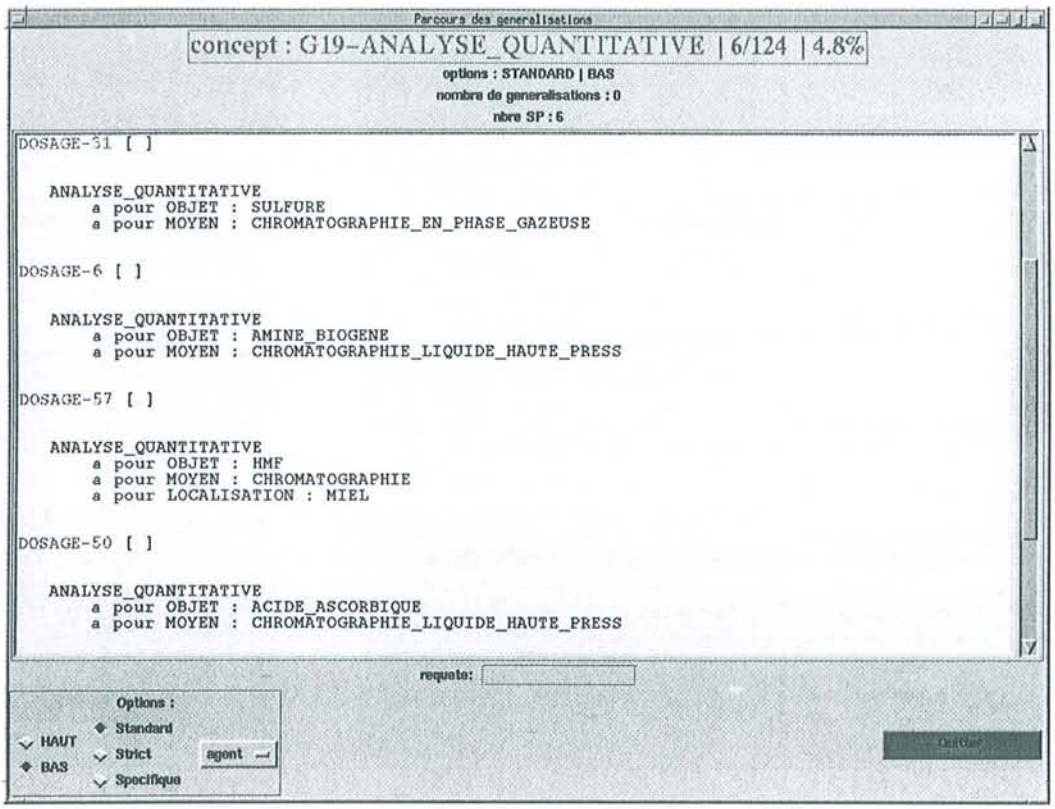

FIG. 6.21 – *Écran de parcours hypertexte des structures prédicatives, faisant apparaître les fils de la généralisation* `G19-ANALYSE_QUANTITATIVE`

les *amines*, qui sont des sortes de *substances chimiques*; et finalement, la *chromatographie*, le moyen, qui est une sorte d'*activité*. Le degré de généralité élevé des rôles *objet* et *moyen* s'explique par le fait que les structures prédicatives extraites avec le prédicat *dosage* sortent des limites de la classe : des termes extérieurs à la classe apparaissent et conduisent à des arguments plus généraux. Par ailleurs, `G19-DOSAGE` montre aussi que les *produits* sont liés au processus de *dosage*, ce qui n'est pas exprimé par la classe.
`G22-DOSAGE` apporte plus de précision sur le processus de dosage : c'est la *chromatographie* qui est utilisé comme activité pour le *dosage* des *substances chimiques*. Cette généralisation couvre les liens L_{10}, L_{12}, L_{14} et L_{16} de la classe (cf. section 6.1).
`G36-DOSAGE` apporte également une plus grande précision en ce qui concerne les *substances chimiques* : chaque *amine* est en fait l'objet du processus de *dosage*, ce qui n'est exprimé que partiellement par le lien L_{15} de la classe.

La mise en commun des toutes ces informations conduit à une interprétation quasi exhaustive de la classe CHROMATOGRAPHIE, plus fine et plus synthétique que celle donnée par les associations : il y a moins d'informations à considérer et les associations ne sont pas limitées à des associations binaires grâce à l'emploi des structures prédicatives. La figure 6.22 montre le résultat que l'on obtient en intégrant ensemble les informations de ces trois structures prédicatives. Le documentaliste caractérise ainsi plus rapidement les

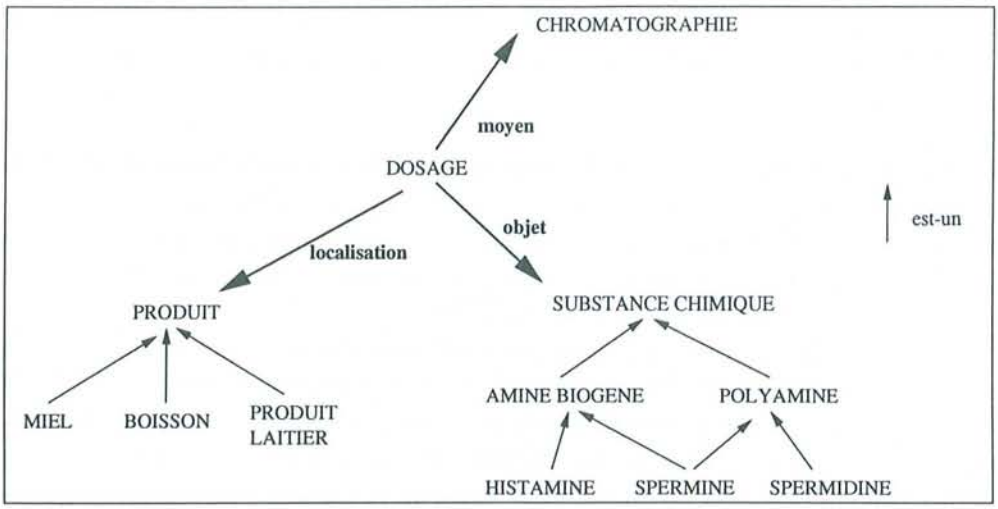

FIG. 6.22 – *Une vue synthétique de la classe* CHROMATOGRAPHIE *construite à partir de trois structures prédicatives*

informations attachées à la classe avec un recours limité aux textes[69].

Quelques problèmes rencontrés lors de l'analyse

Cette expérimentation avec le documentaliste a permis de relever certains problèmes posés par notre approche. Outre le fait que le documentaliste ne connaissait pas les fonctionnalités de notre prototype, nous avons distingué deux types de difficultés :

- l'utilisation d'une hiérarchie de concepts complexe et peu connue de l'expert,
- les insuffisances de notre interface et de notre procédure de généralisation.

La hiérarchie que nous avons utilisée est basée sur le thésaurus AGROVOC, connu par le documentaliste. Toutefois les modifications apportées au thésaurus ont conduit ce dernier à contester plusieurs fois la position de certains concepts. Idéalement, la construction de la hiérarchie devrait être faite de manière consensuelle avec le ou les utilisateurs, ou devrait au moins faire l'objet d'une validation par ceux-ci. Les contraintes matérielles ne nous ont pas permis d'effectuer cette validation, et cela a fortement contribué aux problèmes de compréhension de l'expert. La modélisation et la validation de la hiérarchie sont des activités très coûteuses en temps, qui doivent être prises en compte.

En ce qui concerne l'interface de présentation, nous avons remarqué que la présentation sous format hypertextuel des structures prédicatives est un peu lourde pour parcourir la hiérarchie. Le documentaliste s'est montré peu à l'aise avec la présentation hypertextuelle. Une présentation sous forme graphique de la hiérarchie avec la possibilité de cacher/découvrir la description d'un noeud serait plus adaptée. En effet, dans le cas de l'hypertexte, on ne dispose que d'une vue locale de la hiérarchie. Il faut pouvoir naviguer plus rapidement à travers l'information.

Nous avons enfin constaté que le parcours de la hiérarchie est dans certains cas de figures alourdi par un trop grand nombre de généralisations. En effet, lorsqu'il y a une

69. Il nous semble nécessaire de conserver cette possibilité de retour au document, indispensable pour saisir le contexte complet d'une information.

quantité importante de généralisations, les niveaux hiérarchiques sont multipliés et réduisent la lisibilité de la structure. Ce cas se rencontre notamment avec le prédicat *dosage*, qui apparaît dans les textes avec de nombreux arguments, ce qui conduit à générer de nombreuses structures prédicatives. Les heuristiques que nous avons adoptées pour le processus de généralisation peuvent donc s'avérer insuffisantes. Comme nous l'avons déjà mentionné auparavant, nous ne pouvons demander à l'utilisateur de sélectionner lui-même les informations qu'il juge pertinentes. Une solution qui pourrait être envisagée est d'éliminer certaines généralisations sur la base d'un critère de pertinence qui reste à définir (ce point est discuté dans les perspectives, cf. conclusion finale).
Enfin, le documentaliste aurait aimé avoir accès également à des relations différentes des relations thématiques présentées. Il nous a notamment interrogé sur l'absence de relations de causalité qui peuvent s'avérer utile pour mieux cerner le contenu informationnel. Pour répondre à cette demande, il pourrait être intéressant d'intégrer dans la chaîne de traitement un outil permettant la prise en compte de ce type de relations. Nous pensons notamment au système COATIS [Garcia 96], qui est un outil d'aide à l'acquisition de connaissances causales à partir de textes.

6.5.3 Conclusion

L'expérimentation effectuée est une première évaluation de notre procédure de généralisation appliquée à l'analyse de l'information. Comme nous l'avons remarqué, le fait de ne pas disposer des fonctionnalités de l'interface HENOCH sur notre prototype a limité la portée de cette expérimentation. Il aurait été également préférable de fournir au documentaliste un apprentissage préalable du prototype et de la hiérarchie, ce qui n'a pas été possible.

Il est difficile de conclure définitivement sur la capacité de structuration de notre processus de généralisation, car il n'existe pas de méthode objective d'évaluation. Nous avons remarqué que les généralisations obtenues étaient un peu trop nombreuses pour la classe CHROMATOGRAPHIE, mais que l'ensemble des structures prédicatives obtenues était globalement pertinent.
Mais nous avons surtout montré l'intérêt et la réelle complémentarité des structures prédicatives par rapport aux fonctionnalités existantes de SDOC :

- le contenu informationnel peut être obtenu en grande partie par les structures sans nécessiter un retour aux document du corpus,
- la généralisation des structures prédicatives permet une réelle synthèse d'informations disséminées au sein des documents.

Nous pensons ainsi avoir montré l'intérêt de notre approche pour l'analyse de l'information, qui a été confirmé par le documentaliste.

Conclusion

Notre travail constitue une étude sur la généralisation de structures prédicatives pour l'analyse de l'information. Nous avons voulu montrer l'intérêt d'une méthode de structuration symbolique, la généralisation, pour améliorer une chaîne de traitement dont l'objectif est d'extraire et de synthétiser le contenu informationnel d'un ensemble de textes.

Nous avons scindé la présentation de cette étude en deux grandes parties : nous nous sommes tout d'abord intéressé à la définition d'un processus de généralisation de structures prédicatives, puis nous avons montré comment un tel processus peut s'intégrer dans une méthode d'analyse de l'information afin d'en améliorer l'analyse.

Généralisation de structures prédicatives

Nous avons tout d'abord réalisé une étude critique des travaux en classification conceptuelle, champ dans lequel s'inscrit la généralisation d'objets symboliques. Après avoir distingué deux grands types d'approches, nous nous sommes focalisés sur l'approche de type « Espace de Connaissances » qui permet une meilleure organisation des concepts que l'approche « classique », trop axée sur la prédiction de valeurs. L'approche de type « Espace de Connaissances » a été proposée dans le cadre de plusieurs formalismes de représentation de connaissances : treillis de concepts, représentation par objets, graphes conceptuels. Nous avons mis en évidence les problèmes posés par l'utilisation de ces méthodes pour généraliser des structures prédicatives. Notamment, la génération de toutes les classes (ou concepts) possibles nécessite un travail important de réduction de la hiérarchie qui doit être guidé par l'utilisateur final. Or, dans une perspective d'analyse de l'information, l'utilisateur ne doit pas être surchargé par des tâches annexes à l'analyse elle-même.
Par ailleurs, nous nous sommes intéressés au formalisme des logiques de descriptions, qui présente de nombreux points communs avec les formalismes utilisés pour l'approche de type « Espace de Connaissances », et présente la particularité d'offrir un compromis expressivité/complexité intéressant. Au regard des formalismes cités, nous avons adopté pour notre processus de généralisation une logique de descriptions, CLASSIC, qui nous permet de remédier aux problèmes de complexité posés par les graphes conceptuels ou la représentation par objets.

Nous avons ensuite montré comment la logique de descriptions CLASSIC peut être utilisée pour représenter les structures prédicatives. Nous nous sommes principalement focalisés sur la méthode de représentation. Nous avons alors proposé une méthode de calcul des généralisations des structures prédicatives. Cette méthode adopte une vision logique de la généralisation, en exploitant la notion de plus petit subsumant commun d'un ensemble de concepts. Mais elle prend également en compte la nature des structures manipulées, qui sont issues de langue : nous proposons donc un ensemble d'heuristiques qui permettent de limiter les généralisations calculées en tenant compte d'une hiérarchie de concepts représen-

tant un domaine de connaissances particulier. Le processus est scindé en deux étapes : une première étape permet de calculer des généralisations sur des ensembles de structures prédicatives possédant une tête prédicative donnée ; une deuxième étape calcule des structures prédicatives plus générales à partir de celles obtenues par la première étape. Nous limitons ainsi de manière importante le nombre de généralisations calculées tout en conservant une approche basée sur des opérations logiques, la subsomption et le mécanisme déductif de classification de la logique de descriptions. L'étude de la complexité de notre processus montre qu'elle est comparable aux résultats obtenus par le système de classification conceptuelle COING adoptant une approche de type « Espace de Connaissances ».

La généralisation appliquée à l'analyse de l'information

Dans un deuxième temps, nous avons montré l'intérêt de notre processus de généralisation pour l'analyse de l'information. L'analyse de l'information consiste à collecter l'information contenue dans un corpus de documents afin d'en présenter une vision synthétique à l'utilisateur.

Nous avons tout d'abord étudié les techniques et les méthodes qui permettent d'extraire et de structurer des informations à partir d'une ensemble de textes sous forme électronique. Nous avons montré l'importance d'une approche terminologique pour accéder aux informations des textes, et étudié plusieurs systèmes permettant l'extraction de termes, qui constituent les unités d'informations minimales des textes. Pour structurer ces informations minimales, il est nécessaire d'identifier les relations qui les unissent. Nous avons identifié deux grands types d'approches pour extraire ces relations à partir des textes : une approche à base statistique, qui permet d'obtenir des associations non typées à partir de contextes syntaxiques, et des classes de termes ; une approche symbolique, qui requiert de nombreuses connaissances a priori, et qui permet d'obtenir des structures prédicatives ou structures de connaissances similaires. Parmi ces différentes approches, nous avons étudié plus particulièrement la chaîne de traitement ILIAD qui combine plusieurs techniques (extraction de termes, identification de relations avec une approche statistique) avec l'analyse de l'information comme objectif.

ILIAD permet de construire des classes de termes à partir d'un corpus de documents. Ces classes, composées de termes et de leur associations, constituent le point de départ de l'analyse par un expert. A l'aide d'une carte thématique, et de fonctions d'accès aux titres et aux documents, l'expert effectue l'analyse de l'information contenue dans le corpus.

Dans le but de faciliter le travail de l'expert, nous proposons d'étendre la chaîne de traitement. D'une part, l'utilisation d'une hiérarchie de termes permet de structurer les éléments d'une classe. D'autre part, la collecte des structures prédicatives impliquant les termes d'une classe permet de caractériser plus finement les associations de la classe. Nous montrons que la structuration des structures prédicatives au moyen de notre processus de généralisation permet de synthétiser les informations relatives à une classe.

Dans le dernier chapitre, nous proposons une première évaluation du processus de généralisation dans ce cadre de l'analyse de l'information. Nous nous focalisons sur une classe particulière, CHROMATOGRAPHIE, puis nous décrivons notre prototype mettant en oeuvre le processus de généralisation des structures prédicatives. Nous essayons d'évaluer l'apport de notre proposition en réalisant une expérimentation avec un documentaliste-expert : d'une part avec les outils de la chaîne ILIAD ; d'autre part avec notre prototype. Le résultat montre que notre approche contribue à améliorer l'analyse d'une classe, notamment en permettant une vue synthétique de l'ensemble des informations. Les limitations relevées concernent la modélisation de la hiérarchie de concepts, la nécessité d'une interface de vi-

sualisation plus adaptée, et la présence de généralisations intermédiaires inutiles réduisant la qualité de lecture de la hiérarchie résultante.

Résultats

Nous avons conçu une méthode de généralisation originale qui s'applique sur des structures prédicatives exprimées dans le formalisme des logiques de descriptions. Les caractéristiques de cette méthode sont les suivantes :

- elle est basée sur des mécanismes logiques bien définis (subsomption, classification déductive), contrairement aux approches « classiques » mettant en oeuvre des fonctions d'évaluation d'ordre statistique,
- elle propose un ensemble d'heuristiques permettant de limiter le nombre de généralisations de manière automatique, contrairement aux approches de type « Espace de Connaissances », qui nécessitent une intervention forte de l'utilisateur,
- elle possède une complexité similaire aux approches de type « Espace de Connaissances »,
- elle s'applique à des structures prédicatives. Elle est donc moins générale que les autres approches qui s'appliquent à des concepts quelconques[70], mais permet une meilleure prise en compte de la spécificité des structures prédicatives, notamment pour la définition des heuristiques.

D'autre part, l'application de notre processus de généralisation à l'analyse de l'information nous a permis :

- de montrer l'importance d'une approche terminologique pour l'extraction d'information, et de montrer la faisabilité d'une méthode d'analyse de l'information,
- de proposer une amélioration de la chaîne de traitement ILIAD pour un accès plus rapide au contenu informationnel des textes,
- d'effectuer une première évaluation de notre processus de généralisation avec un expert documentaliste, qui montre des résultats encourageants.

Généralité de notre approche

La généralité de notre approche peut être jugée selon deux critères principaux : d'une part, quelles sont les modifications nécessaires pour prendre en compte un domaine de connaissances différent de celui que nous avons choisi pour notre expérimentation? D'autre part, notre processus de généralisation peut-il faire l'objet d'applications différentes de l'analyse de l'information?

Notre méthode de généralisation n'est pas restreinte à un domaine de connaissances particulier. Toutes les fonctions définies sont totalement indépendantes des données liées au domaine. Par contre, il est bien sûr nécessaire de fournir ces données, qui sont : un ensemble de termes (y compris les prédicats) du domaine ; une organisation de ces termes en hiérarchie ; un ensemble de structures prédicatives relatives au domaine. Par ailleurs, pour utiliser la chaîne de traitement ILIAD, il est nécessaire de fournir des informations d'ordre linguistique sur les termes. Le passage d'un domaine de connaissances à un autre n'est

70. Ces concepts sont cependant restreints par le pouvoir expressif du langage utilisé.

donc pas quelque chose de négligeable. Il s'agit d'une activité importante de modélisation indispensable pour prendre en compte la spécificité du domaine considéré.

Notre processus de généralisation est particulièrement adapté aux structures prédicatives. Il n'est cependant pas exclu de l'utiliser pour les concepts d'une base de connaissances, à la condition de respecter les restrictions de représentations imposées par le processus. Nous pouvons envisager également de l'appliquer à la recherche d'informations à la manière de Carpineto et Romano (cf. section 2.4), en considérant les structures prédicatives à la place des descripteurs pour la recherche des documents.

Perspectives

Les perspectives autour de l'analyse de l'information et de notre processus de généralisation sont multiples. Nous discutons des trois idées principales que nous souhaiterions développer en priorité.

En premier lieu, il nous semble important de définir et proposer une méthode d'extraction de structures prédicatives « légère », focalisée sur les groupes nominaux, afin de compléter et finaliser la chaîne de traitement pour l'analyse de l'information. La collecte des structures prédicatives est à présent manuelle et constitue un frein important à l'utilisation réelle de notre méthode. La méthode d'extraction devra être moins coûteuse que celles des systèmes RECIT et PAPINS que nous avons étudiés : d'une part en restreignant les structures syntaxiques à analyser ; d'autre part en se focalisant sur les arguments essentiels, au détriment des arguments secondaires, moins important en terme d'information. Un travail important sur cette problématique est en cours de réalisation : J. Royauté [Royauté 98] a étudié très finement les groupes nominaux à tête prédicative dans une perspective d'analyse de l'information. Dans la troisième partie de sa thèse, il propose une acquisition assistée de prédicats ainsi qu'une méthode de recherche des structures argumentales dans les groupes nominaux complexes. Ce travail permettra sans doute d'aboutir à une méthode d'extraction nécessitant moins de connaissances a priori que les méthodes d'extraction dont nous avons discuté.

La deuxième point concerne notre prototype, qui nécessite d'être amélioré au niveau de l'interface utilisateur. Notre travail a été peu important sur cet aspect, qui est cependant primordial pour aboutir à une méthode effective. Il nécessite une réflexion approfondie, en lien avec les utilisateurs potentiels. L'intégration des fonctionnalité de HENOCH, l'interface de visualisation de SDOC, et la définition de méthodes de parcours de la hiérarchie plus ergonomiques, permettraient d'obtenir une efficacité accrue lors de l'analyse des textes.

Enfin, nous avons remarqué que notre processus de généralisation, malgré les heuristiques définies, génère des structures prédicatives qui réduisent la lisibilité de la structure hiérarchique. Nous pensons qu'une solution possible consiste à définir des critères de sélection permettant d'éliminer automatiquement les structures prédicatives qui sont jugées les moins pertinentes. Cela revient à la solution proposée par Bournaud, sans toutefois faire intervenir l'utilisateur dans le processus. Les critères proposés par Bournaud (cf. section 2.5.3) peuvent constituer un point de départ intéressant. Nous envisageons cependant de définir un plus grand nombre de paramètres pouvant représenter les critères intuitifs suivants :

- une généralisation n'est pas trop générale : les termes impliqués ne doivent pas être situés trop haut dans la hiérarchie, et le nombre de relations doit être le plus élevé

possible,

- une généralisation n'est pas trop spécifique : elle doit « couvrir » un maximum de structures prédicatives,
- une généralisation doit être homogène : les termes impliqués ne doivent pas être situés à des niveaux trop différents relativement les uns aux autres dans la hiérarchie,
- une généralisation doit synthétiser plusieurs informations : elle doit donc posséder un maximum de descendants directs.

Chacun de ces critères constitue un indice de la pertinence d'une généralisation. Il faut toutefois être capable de trouver une formulation précise qui puisse rendre compte de ces indices, de manière logique ou numérique.

Nous pensons que l'analyse de l'information est un domaine de recherche qui est appelé à se développer et permettra, dans le contexte de l'Internet et de l'importance toujours plus grande du support électronique des documents, de mieux maîtriser les flux d'informations. Notre travail est une contribution qui, nous l'espérons, permettra de disposer à terme d'outils performants pour l'analyse de contenu, essentiels pour assister les utilisateurs dans l'exploration et l'analyse de documents électroniques.

Bibliographie

[Agarwal 94] R. Agarwal. (almost) automatic semantic feature extraction from technical text. *Proceedings of the ARPA Human Language Technology Workshop*, New Jersey, 1994.

[Agarwal 95] R. Agarwal. *Semantic Feature Extraction from Technical Texts with Limited Human Intervention.* Thèse de Doctorat, Mississippi State University, 1995.

[Assadi 96] H. Assadi et D. Bourigault. Acquisition et modélisation de connaissances à partir de textes : outils informatiques et éléments méthodologiques. *Actes RFIA'96, Rennes, 15-18 janvier 96, volume 1*, pages 505–514, 1996.

[Assadi 97] H. Assadi. Une méthode et des outils pour la construction d'un modèle du domaine à partir de textes. Application à la consultation d'une documentation technique. *Actes des journées Ingénierie des Connaissances et Apprentissage Automatique, JICAA'97, Roscoff, France, 20-22 Mai, 1997*, 1997.

[Aussenac-Gilles 95] N. Aussenac-Gilles, D. Bourigault, A. Condamines et C. Gros. How can knowledge acquisition benefit from terminology? *Proceedings of the 9th Knowledge Acquisition Workshop*, Banff, CA, 1995.

[Baader 94] F. Baader, B. Hollunder, B. Nebel, H.J. Profitlich et E. Franconi. A empirical analysis of optimization techniques for terminological representation systems. *Journal of Applied Intelligence*, 4(2) :109–132, 1994.

[Bachimont 95] B. Bachimont. Ontologie régionale et terminologie : quelques remarques méthodologiques et critiques. In Otman [Otman 95], pages 67–86.

[Barrière 96] C. Barrière et F. Popowich. Concept clustering and knowledge integration from a children's dictionnary. *Proceedings of 16th International Conference on Computational Linguistics (COLING-96)*, pages 65–70, 1996.

[Beck 94] H.W. Beck, T. Anwar et S.B. Navathe. A conceptual clustering algorithm for database schema design. *IEEE Transactions on Knowledge and Data Engineering*, pages 396–411, june 1994.

[Biebow 91] B. Biebow et S.Szulman. Interprétation de groupes nominaux complexes dans un réseau sémantique à la kl-one. *Actes du 8ème congrès RFIA (RFIA'91)*, 1991.

[Bisson 92] G. Bisson. Conceptual clustering in a first order logic representation. B. Neumann, editor, *Proceedings of the tenth european conference on artificial intelligence*, pages 458–462, Vienna, Austria, 1992. John Wiley & Sons.

[Boguraev 96] B. Boguraev et J. Pustejovsky, editors. *Corpus Processing for Lexical Acquisition.* The MIT Press, 1996.

[Borgida 89] A. Borgida, R. Brachman, D. McGuinness et L. Resnick. Classic : a structural data model for objects. *SIGMOD-89*, pages 58–67, 1989.

[Borgida 96] A. Borgida. On the relative expressiveness of description logics and predicate logics. *Artificial Intelligence*, 82(1-2) :353–367, 1996.

[Bouaud 97] J. Bouaud, B. Habert, A. Nazarenko et P. Zweigenbaum. Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles. *Actes des journées Ingénierie des Connaissances et Apprentissage Automatique, JICAA'97, Roscoff, France, 20-22 Mai, 1997*, 1997.

[Bourigault 94a] D. Bourigault. *LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition de connaissances à partir de textes.* Thèse de Doctorat, Ecole des Hautes Etudes en Sciences Sociales, juin 1994.

[Bourigault 94b] D. Bourigault et P. Lépine. Méthodologie d'utilisation de LEXTER pour l'acquisition des connaissances à partir de textes. In JAC [JAC 94], pages F1–F13.

[Bourigault 95] D. Bourigault et A. Condamines. Réflexions sur le concept de base de connaissances terminologiques. *Actes des 5èmes journées nationales P.R.C.-G.D.R. Intelligence Artificielle*, pages 425–444, Nancy, 1995. Teknea.

[Bournaud 96] I. Bournaud. *Regroupement conceptuel pour l'organisation des connaissances.* Thèse de Doctorat, Thèse de l'université Paris 6, 1996.

[Brachman 78] R. Brachman. *A structural paradigm for representing knowledge.* Bolt Beranek and Newman Inc, 1978.

[Brachman 91] R.J. Brachman, D.L. Mc Guinness, P.F. Patel-Schneider, L.A. Resnick et A. Borgida. *Principles of Semantics Networks. Exploration in the Representation of Knowledge*, chapitre Living with CLASSIC : When and How Use a KL-ONE Language, pages 401–456. Morgan Kaufmann, 1991.

[Brill 93] E. Brill. *A Corpus-Based Approach to Language Learning.* Thèse de Doctorat, University of Pennsylvania, 1993.

[Capponi 97a] N. Capponi. Use of description logics for shallow information analysis from texts. *Proceedings Description Logic Workshop (DL' 97), 27-29 september 1997*, Gif-sur-Yvette, France, 1997.

[Capponi 97b] N. Capponi et Y. Toussaint. The ILIAD Project : Analysing Information Using Informetrics Techniques and Natural Language Processing. *Actes du Third DELOS Workshop on Cross-Language Information Retrieval, 5-7 March 1997*, Zurich, Switzerland, 1997. ERCIM.

[Capponi 98a] N. Capponi. Analyse de l'information contenue dans des textes scientifiques avec une logique de descriptions. *Actes du 11ème congrès RFIA (RFIA'98), 20-22 janvier 1998*, Clermont-Ferrand, France, 1998.

[Capponi 98b] N. Capponi et Y. Toussaint. Interprétation de classes de termes par généralisation de structures prédicat-argument. *Actes du colloque Ingénierie des Connaissances (IC'98), 13-15 mai 1998*, pages 41–50, Pont-à-Mousson, France, 1998.

[Carpineto 93] C. Carpineto et G. Romano. Galois : An order-theoretic approach to conceptual clustering. *Proceedings of the Tenth International Conference on Machine Learning*, pages 33–40, 1993.

[Carpineto 96] C. Carpineto et G. Romano. A lattice conceptual clustering system and its application to browsing retrieval. *Machine Learning*, 24 :95–122, 1996.

[Celeux 95] G. Celeux, E. Diday, G. Govaert, Y. Lechevallier et H. Ralambondrainy. *Classification Automatique de Données*. Eyrolles, Paris, 1995.

[Cercone 87] N. Cercone et G. Mac Colla. *The Knowledge Frontier : Essays in the Representation of Knowlege*, chapitre What is Knowledge Representation? Springer Verlag, 1987.

[Charlet 94] J. Charlet, B. Bachimont, J. Bouaud et P. Zweigenbaum. Ontologie et réutilisabilité : expérience et discussion. In JAC [JAC 94], pages C1–C14.

[Church 90] K.W. Church et P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1) :22–29, 1990.

[Cohen 92] W.W. Cohen, A. Borgida et H. Hirsh. Computing least common subsumers in description logics. *Proceedings of the Tenth National Conference on Artificial Intelligence*, San Jose, California, 1992. MIT Press.

[Cohen 94a] W. Cohen et H. Hirsh. The learnability of description logics with equality constraints. *Machine Learning*, 17(2-3) :169–199, 1994.

[Cohen 94b] W.W. Cohen et H. Hirsh. Learning the CLASSIC Description Logic : Theoretical and Experimental Results. *Proceedings of the Fouth International Conference on Principles of Knowledge Representation and Reasoning (KR'94)*, 1994.

[Condamines 92] A. Condamines. Aide à l'acquisition de connaissances par la spécification de la terminologie d'un domaine de spécialité. *Actes Journées d'Acquisition de Connaissances (JAC'92)*, 1992.

[Condamines 97] A. Condamines et J. Rebeyrolle. Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. *Actes de JICAA'97*, pages 191–206, 1997.

[Copeck 97] T. Copeck, K. Barker, S. Delisle, S. Szpakowicz et J.F. Delannoy. What is a technical text? *Language Sciences*, 19(4) :391–424, 1997.

[Coupey 97] P. Coupey et S. Salotti. Une logique de descriptions comme cadre formel d'un système de raisonnement à partir de cas. *Revue d'intelligence artificielle*, 11(2) :127–177, 1997.

[Cowie 96] J. Cowie et W. Lehnert. Information extraction. *Communications of the ACM*, 39(1) :80–91, janvier 1996.

[Croft 92] W.B. Croft. *Text-Based Intelligent Systems : Current Research and Practice in Information Extraction and Retrieval*, chapitre Text retrevial and inference. Lawrence Erlbaum Ass., Hillsdale, NJ, 1992.

[Cruse 86] D.A. Cruse. *Lexical Semantics.* Cambridge University Press, 1986.

[Daille 94] B. Daille. *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques.* Thèse de doctorat, Université de Paris VII, 1994.

[Daille 96] B. Daille, B. Habert, C. Jacquemin et J. Royauté. Empirical observation of term variations and principle for their description. *Terminology*, 3(2) :197–257, 1996.

[Decaestecker 93] C. Decaestecker. Apprentissage et outils statistiques en classification conceptuelle incrémentale. *Revue d'intelligence artificielle*, 7(1) :33–71, 1993.

[Delisle 96] S. Delisle, K. Barker, T. Copeck et S. Szpakowicz. Interactive semantic analysis of technical texts. *Computational Intelligence*, 12(2) :273–306, 1996.

[Donini 97] F.M. Donini, M. Lenzerini, D. Nardi et W. Nutt. The complexity of concept languages. *Information and Computation*, 134(1) :1–58, 1997.

[Doyle 91] J. Doyle et R. Patil. Two theses of knowledge representation : language restrictions, taxonomic classification, and the utility of representation services. *Artificial Intelligence*, 48(3) :261–298, 1991.

[Ducournau 96] R. Ducournau. Des langages à objets aux logiques terminologiques : les systèmes classificatoires. Rapport de recherche no. 96-030, LIRMM, Montpellier, 1996.

[Dumas 96] L. Dumas, A. Plante et P. Plante. Nomino ; version 1.0. Rapport, Centre ATO, Département de linguistique, Université du Québec à Montréal, 1996.

[Euzenat 94] J. Euzenat. Classification dans les représentations par objets : produits des systèmes classificatoires. *Actes du 9ème congrès Reconnaissance des Formes et Intelligence Artificielle, RFIA '94*, volume 2, pages 185–196, 1994.

[Felber 87] Helmut Felber. *Manuel de terminologie.* Infoterm/Unesco, 1987.

[Fillmore 68] C. Fillmore. *Universals in linguistic theory*, chapitre The case for case. Holt, Rinehart and Winston, New York, 1968. E. Bach & R. Harms eds.

[Fisher 87] D. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2 :139–172, 1987.

[Forster 94] P. Forster. Lexical semantics, description logics and natural language systems. F. Baader, M. Lenzerini, W. Nutt et P.F. Patel-Schneider, editors, *Proceedings of International Workshop on Description Logics DL'94*, pages 84–86. DFKI, 1994.

[Franconi 94] E. Franconi. Description logics for natural language processing. *Working Notes of the 1994 AAAI Fall Symposium on « Knowledge Representation for Natural Language Processing in Implemented Systems », New Orleans.*, 1994.

[Fuchs 93] C. Fuchs, A. Lacheret-Dujour et B. Victorri. *Linguistique et Traitements Automatiques des Langues.* Hachette, Paris, 1993.

[Garcia 96] D. Garcia. Coatis, un outil d'aide à l'acquisition de connaissances causales exprimées dans les textes. P. Bouffard et A. Kharrat, coordi-

nateurs, *Actes du premier colloque étudiant de linguistique informatique de Montréal (CLIM-96), 8-10 juin 1996*, pages 96–103, Montréal, Canada, 1996. CLIM, Département de linguistique et de traduction, Université de Montréal.

[Gey 94] O. Gey. Saturation et généralisation de graphes conceptuels. *Actes de JFA '94*, 1994.

[Godin 95] R. Godin, G. Mineau, R. Missaoui et H. Mili. Méthodes de classification conceptuelle basées sur les treillis de galois et applications. *Revue d'intelligence artificielle*, 9 :105–137, 1995.

[Gouadec 94] D. Gouadec. *Données et informations terminologiques et terminographiques : natures et valeurs*, volume 1, série *Terminoguides*. La Maison du Dictionnaire, 1994.

[Grefenstette 93] G. Grefenstette. Evaluation techniques for automatic semantic extraction : Comparing syntactical and window based approaches. *Workshop on Acquisition of Lexical Knowledge form Text. SIGLEX/ACL*, Colombus, USA, juin 1993.

[Grefenstette 94] G. Grefenstette. Corpus-derived first, second and third-order word affinities. Rapport no. MLTT-09, Rank Xerox Research Center, Grenoble Laboratory, 1994.

[Grivel 95a] L. Grivel et C. François. *Les sciences de l'information. Bibliométrie. Scientométrie. Infométrie.*, chapitre Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique, pages 81–113. Presses universitaires de Rennes, 1995.

[Grivel 95b] L. Grivel, P. Mutschke et X. Polanco. Thematic mapping on bibliographic databases by cluster analysis : a description of the sdoc environment with solis. *Journal of Knowledge Organization*, 22(2) :70–77, 1995.

[Grivel 97] L. Grivel et C. François. Deux éléments de la plate-forme infométrique de l'INIST : NEURODOC et HENOCH. *Séminaire de l'ADEST, 9 décembre 1997, Univ. Pierre Mendès-France, Grenoble*, 1997. http ://melpomene.upmf-grenoble.fr/adest/seminaires/francois.htm.

[Guha 90] R.V. Guha et D.B. Lenat. Cyc : a midterm report. *AI Magazine*, pages 32–59, 1990.

[Habert 96a] B. Habert, P. Barbaud, F. Dupuis et C. Jacquemin. Simplifier des arbres d'analyse pour dégager des comportements syntactico-sémantiques des formes d'un corpus. *Cahiers de Grammaire*, 1996.

[Habert 96b] B. Habert et A. Nazarenko. La syntaxe comme marche-pied de l'acquisition des connaissances : bilan critique d'une expérience. *Actes Journées Acquisition des Connaissances (JAC'96)*, 1996.

[Habert 97] B. Habert, A. Nazarenko et A. Salem. *Les linguistiques de corpus.* Armand Colin, 1997.

[Harris 89] Z. Harris, M. Gottfried, T. Ryckman, P. Mattick Jr., A. Daladier, T.N. Harris et S. Harris. *The Form of Information in Science*, volume 104, série *Boston Studies in the Philosophy of Science.* Kluwer Academic Publishers, 1989.

[Haussler 89] D. Haussler. Learning conjunctive concepts in structural domains. *Machine Learning*, 4 :7–40, 1989.

[Heinsohn 92] J. Heinsohn, D. Kudenko, B. Nebel et H.-J. Profitlich. An empirical analysis of terminological representation systems. Rapport no. RR-92-16, Deutsches Forschungszentrum fur Künstliche Intelligenz GmbH (DFKI), 1992.

[Herviou 95] M.L. Herviou. Applications d'extraction de connaissances à EDF-DER. *Journées IA 95, Montpellier, France*, 1995.

[JAC 94] PRC-GDR IA, CNRS. *Actes des Cinquièmes Journées Acquisition des Connaissances, Strasbourg, 21-23 Mars 1994*, 1994.

[Jackendoff 90] R. Jackendoff. *Semantic structures.* MIT Press, Cambridge, MA, 1990.

[Jacquemin 95] C. Jacquemin. A symbolic and surgical acquisition of terms through variation. *Proceedings of the Workshop on "New approaches to learning for NLP" at the 14th International Joint Conference on Artificial Intelligence (IJCAI'95), Montréal*, 1995.

[Jacquemin 97] C. Jacquemin. Variation terminologique : Reconnaissance et acquisition automatique de termes et de leur variantes en corpus. Habilitation à diriger des recherches, Université de Nantes, Faculté des Sciences, Nantes, 1997.

[Jansen-Winkeln 91] R. Jansen-Winkeln, A. Ndiaye et N. Reithinger. FSS-WASTL : Interactive Knowledge Acquisition for a Semantic Lexicon. *Proceedings of 2nd Congress of the Italian Association for Artificial Intelligence,AI*IA, Palermo, Italy, october 1991*, volume 549, série *Lecture Notes in Artificial Intelligence*, 1991.

[Ketterlin 95] A. Ketterlin. *Découverte de Concepts Structurés dans les Bases de Données.* Thèse de Doctorat, Département d'Informatique, Université Louis Pasteur, France, 1995.

[Kietz 94] J. Kietz et K. Morik. A polynomial approach to the constructive induction of structural knowledge. *Machine Learning*, 14(2) :193–217, 1994.

[Leclère 96] M. Leclère. Définition de types dans le modèle des graphes conceptuels. *Actes RFIA'96, Rennes, 15-18 janvier 96, volume 1*, pages 486–493, 1996.

[Lerat 90] P. Lerat. L'hyperonymie dans la structuration des terminologies. *Langages*, (98) :79–86, juin 1990.

[Lerat 95] P. Lerat. *Les langues spécialisées.* coll. Linguistique Nouvelle. Presses Universitaires de France, 1995.

[Lindberg 93] D.A. Lindberg, B.L. Humphreys et A.T. McRay. The unified medical language system. *Methods of Information in Medicine*, 32(4) :281–291, 1993.

[Lyons 77] J. Lyons. *Semantics.* Cambridge University Press, 1977.

[MacGregor 94] R. MacGregor et D. Brill. Recognition algorithms for the LOOM classifier. *Proceedings of the Twelfth National Conference on Artificial Intelligence, (AAAI 94)*, pages 213–220, 1994.

[Mariño 93] O. Mariño. *Raisonnement classificatoire dans une représentation à objets multi-points de vue.* Thèse de Doctorat, Université Joseph Fourier, Grenoble (FR), 1993. ftp://ftp.inrialpes.fr/pub/sherpa/theses/marino.ps.gz.

[Mauldin 91] M. Mauldin. Retrieval Performance in FERRET : A Conceptual Information Retrieval System. *Proceedings of the 14th International Conference on research and Development in Information Retrieval, Chicago, october 1991, ACM SIGIR*, 1991.

[Meyer 91] I. Meyer, L. Bowker et K. Eck. Constructing a knowledge-based term bank : fundamentals and implications. *Actes de International Symposium on Terminology and Documentation in Specialized Communication, Hull, Canada, 7-8 octobre 1991*, 1991.

[Meyer 94] I. Meyer. Helping terminologists do knowledge engineering : Some linguistic strategies and computers aids. *Actualité Terminologique*, pages 6–10, décembre 1994.

[Michalski 83] R.S. Michalski et R.E. Stepp. *Machine learning : An artificial intelligence approach, Volume 1*, chapitre Learning from observation : Conceptual clustering, pages 331–363. Morgan Kaufmann, 1983.

[Mikheev 95] A. Mikheev et S. Finch. Toward a workbench for acquisition of domain knowledge from natural language. *Proceedings of seventh conference of the European Chapter of the Association for Computational Linguistics, Dublin , Ireland, March 1995*, pages 194–201, 1995.

[Miller 93] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross et K. Miller. Five papers on wordnet. Rapport, Cognitive Science Laboratory, Princeton University, 221 Nassau St., Princeton, NJ 08542, 1993.

[Mineau 90] G. Mineau et G. Godin. La classification symbolique : une approche non-subjective. *Actes des 5èmes Journées Françaises sur l'Apprentissage (JFA)*, pages 169–189, Lanion, France, 1990. CNET.

[Mineau 95] G.W. Mineau et R. Godin. Automatic structuring of knowledge bases by conceptual clustering. *IEEE Transactions in Knowledge and Data Engineering*, 7(5), octobre 1995.

[Mitchell 82] T. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2) :203–226, 1982.

[Muggleton 92] S. Muggleton. *Inductive logic programming.* The A.P.I.C. Series. Academic Press, 1992.

[Mugnier 96] M.-L. Mugnier et M. Chein. Représenter des connaissances et raisonner avec des graphes. *Revue d'intelligence artificielle*, 10(1) :7–56, 1996.

[Muller 97] C. Muller, X. Polanco, J. Royauté et Y. Toussaint. Acquisition et structuration de connaissances en corpus : éléments méthodologiques. Rapport no. RR-3198, INRIA, juin 1997.

[Munday 95] C. Munday, T. Cross, J. Daengdej et D. Lukose. *CGKEE : Conceptual Graph Knowledge Engineering Environment User and System Manual, version 1.0.* Distributed Artificial Intelligence Center, University of New England, Armidale, Australia, août 1995.

[Napoli 96] A. Napoli. Classification et organisation hiérarchique des connaissances. Aspects de la classification, rapport 96-R-072, CRIN, Nancy, 1996.

[Napoli 97] A. Napoli. Une introduction aux logiques de descriptions. Rapport de recherche no. RR-3314, INRIA Lorraine, décembre 1997.

[Nebel 90a] B. Nebel. *Reasoning and Revision in Hybrid Representation Systems.* Lecture Notes in Artificial Intelligence. Springer Verlag, 1990.

[Nebel 90b] B. Nebel. *Reasoning and Revision in Hybrid Representation Systems*, chapitre Representation, Reasoning and Revision - The Idea. In *Lecture Notes in Artificial Intelligence* [Nebel 90a], 1990.

[Nobecourt 98] J. Nobecourt. Représenter la notion de propriété dans les graphes conceptuels et les logiques de descriptions. *Actes du colloque Ingénierie des Connaissances, IC'98, 13-15 mai 1998, Pont-à-Mousson*, 1998.

[Nutter 89] J.T. Nutter. A lexical relation hierarchy. Rapport no. 89-5, Computer Science Department, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1989.

[Orlac 94] I. Orlac et J. Rebeyrolle. Elaboration et validation d'une méthode de recueil des connaissances terminologiques propres à un domaine. Rapport de stage, DESS Sciences Cognitives et Interaction Homme/Machine. Université Toulouse-Le Mirail, 1994.

[Otman 93] G. Otman. La représentation de la relation d'hypéronymie/hyponymie. *ICO*, 5(3), 1993.

[Otman 94] G. Otman. La modélisation des unités terminologiques sous la forme de réseaux sémantico-terminologiques. *Actes Premier Colloque Jeunes Chercheurs en Sciences Cognitives*, La Motte d'Aveillans, mars 1994.

[Otman 95] G. Otman, coordinateur. *Premières Rencontres Terminologie et Intelligence Artificielle, numéro spécial de la Banque des mots 7/1995.* Conseil international de la langue francaise, 1995.

[Oueslati 96] R. Oueslati. Acquisition de termes et de schémas linguistiques pour l'identification de concepts. *Actes du Colloque Informatique & Langue Naturelle (I.L.N.'96), 9-10 octobre 1996, Nantes*, pages 457–463, 1996.

[Ounis 95] I. Ounis. Une dénotation pour les graphes conceptuels : comparaison avec les logiques terminologiques en recherche d'information. *Actes INFORSID'95*, 1995.

[Pichon 97] R. Pichon et P. Sébillot. Acquisition automatique d'informations lexicales à partir de corpus. Rapport no. 3321, INRIA, décembre 1997.

[Pugeault 95a] F. Pugeault. *Extraction dans les textes de connaissances structurées : une méthode fondée sur la sémantique lexicale linguistique.* Thèse de doctorat, Université Paul Sabatier, IRIT, Toulouse, octobre 1995. n. 2153.

[Pugeault 95b] F. Pugeault et M-G. Monteil. Une étude pour l'extraction d'index structurés à la direction des études et recherche d'EDF. *Journées Internationales IA 95, session Génie Linguistique*, Montpellier, juin 1995.

[Pugeault 96] F. Pugeault et G. Lapalme. Vers une génération automatique de synthèses de textes techniques à partir de formes prédicat-arguments. *Actes d'ILN'96*, Nantes, octobre 1996.

[Rassinoux 94] A.M. Rassinoux. *Extraction et représentation de la connaissance tirée de textes médicaux.* Thèse de Doctorat, Faculté des sciences de l'université de Genève, 1994.

[Rastier 95] F. Rastier. Le terme : entre ontologie et linguistique. In Otman [Otman 95], pages 35–65.

[Rector 96] A.L. Rector, J.E. Rogers et P. Pole. The GALEN High Level Ontology. *Medical Informatics in Europe (MIE'96), Copenhague*, 1996.

[Resnick 95] L.A. Resnick, A. Borgida, R.J. Brachman, C.L. Isbell, P.F. Patel-Schneider D.L. Mc Guinness et K.C. Zalondek. Classic description and reference manual for the common lisp implementation (version 2.3). Rapport, AT&T Bell Laboratories, 1995.

[Reynaud 94] C. Reynaud et F. Tort. Connaissances du domaine d'un sbc et ontologies : discussion. In JAC [JAC 94], pages B1–B13.

[Royauté 98] J. Royauté. *Les groupes nominaux complexes et leurs propriétés : Application à l'analyse de l'information.* Thèse de Doctorat, Université Henri Poincaré, Nancy 1, 1998. A paraître.

[Sager 90] J.C. Sager. *A Practical Course In Terminology Processing.* John Benjamins Publishing Company, 1990.

[Saint-Dizier 95] P. Saint-Dizier et E. Viegas. *Computational lexical semantics*, chapitre An introduction to lexical semantics from a linguistic and a psycholinguistic perspective, pages 1–29. Studies in Natural Language Processing. Cambridge University Press, 1995.

[Saitta 96] L. Saitta. Representation change in machine learning. *AI Communications*, 9 :14–20, 1996.

[Salton 94] G. Salton, L. Allan et C. Buckley. Automatic structuring and retrieval of large text files. *Communications of the ACM*, 37(2) :97–108, février 1994.

[Saporta 90] G. Saporta. *Probabilités, Analyse de Données et Statistique.* Ed. Tecnip, Paris, 1990.

[Schmidt-Schauß89] M. Schmidt-Schauß. Subsumption in kl-one is undecidable. *Proceedings of First International Conference on Principles of Knowledge Representation and Reasoning (KR'89)*, pages 421–431, Toronto, Ontario, May 1989.

[Simon 98] A. Simon et A. Napoli. Treillis de galois et représentation par objets pour la fouille de données. *Actes du colloque Ingénierie des Connaissances, IC'98, 13-15 mai 1998, Pont-à-Mousson*, 1998.

[Skuce 91] D. Skuce et I. Meyer. Terminology and knowledge acquisition : exploring a symbiotic relationship. *Proceedings of the 6th Banff Knowledge Acquisition for Knowledge-based Systems Workshop*, pages 29/1–29/21, 1991.

[Sowa 84] J.F. Sowa. *Conceptual Structures : Information Processing in Mind and Machine.* Addison-Wesley, 1984.

[Sowa 91a] J. Sowa. *Principles of Semantics Networks. Exploration in the Representation of Knowledge*, chapitre Toward the Expressive Power of Natural Language. In PSN [Sowa 91b], 1991.

[Sowa 91b] J. Sowa, editor. *Principles of Semantics Networks. Exploration in the Representation of Knowledge.* Morgan Kaufmann, 1991.

[Stephens 94] Charlotte S. Stephens. The nature of information technology research : a seven year analysis. *Journal of Computer System Information Systems*, 34(4) :67–76, Summer 1994.

[Thompson 91] K. Thompson et P. Langley. *Concept Formation : Knowledge and Experience in Unsupervised Learning*, chapitre Concept formation in structured domains, pages 127–161. Morgan Haufmann, San Mateo, California, 1991.

[Toussaint 96] Y. Toussaint. Combining informetrics and linguistics in order to analyse large documentary databases. *Proceedings of KBCS'96*, 1996.

[Toussaint 97] Y. Toussaint, J. Royaute, C. Muller et X. Polanco. Analyse linguistique et infométrique pour l'acquisition et la structuration des connaissances. *Terminologie et Intelligence Artificielle (TIA'97), 3-4 avril 1997, Université Toulouse-Le Mirail, France*, 1997.

[Toussaint 98] Y. Toussaint, F. Namer, B. Daille, C. Jacquemin, J. Royaute et N. Hathout. Une approche linguistique et statistique pour l'analyse de l'information en corpus. *Conférence Traitement Automatique des Langues Naturelles (TALN'98), 10-12 juin 1998*, Paris, 1998.

[Van-Bakel 96] B. Van-Bakel, R. T. Boon, N. J. Mars, J. Nijhuis, E. Oltmans et P. Van der Vet. Condorcet annual report. Rapport no. UT-KBS-96-12, Knowledge-based Systems Group, University of Twente, The Netherlands, septembre 1996.

[Ventos 95] V. Ventos, P. Brézellec, P. Coupey et H. Soldano. C-classic : un langage de descriptions « PAC-learnable ». *Actes JAVA'95*, 1995.

[Viallet 94] F. Viallet, J. Garraud et G. Otman. Administration de données et terminologie. CNRS-URA 1576, Centre de terminologie et de néologie, Université de Paris-Nord, Villetaneuse, non paru, 1994.

[Wille 84] R. Wille. *Ordered Sets*, chapitre Restructuring lattice theory : an approach based on hierarchies of concepts. D. Reidel, I. Rival édition, 1984.

[Winston 87] M.E. Winston, R. Chaffin et D. Herrmann. A taxonomy of part-whole relations. *Cognitive Science*, 11 :417–444, 1987.

[Woods 91] W.A. Woods. *Principles of Semantics Networks. Exploration in the Representation of Knowledge*, chapitre Understanding Subsumption and Taxonomy : A Framework for Progress, pages 45–94. In Sowa [Sowa 91b], 1991.

[Woods 92] W.A. Woods et J.G. Schmolze. The kl-one family. *Computers Math. Applic.*, 23(2-5) :133–177, 1992.

[Zweigenbaum 94] P. Zweigenbaum, B. Bachimont, J. Bouaud, J. Charlet et J.F. Boisvieux. Structuration et acquisition d'une ontologie pour la compréhension du langage médical. Rapport, DIAM, 1994.

[Zweigenbaum 97] P. Zweigenbaum et J. Bouaud. Construction d'une représentation sémantique en graphes conceptuels à partir d'une analyse LFG. *Conférence Traitement Automatique des Langues Naturelles (TALN'97)*, Grenoble, juin 1997.

Résumé

Nous présentons dans ce mémoire une méthode originale de structuration de structures prédicatives par généralisation, que nous appliquons ensuite à l'analyse du contenu informationnel de textes scientifiques.
Dans une première partie, nous présentons notre méthode de généralisation de structures prédicatives. Une étude des méthodes de généralisation existantes, principalement issues du domaine de la classification conceptuelle, montre qu'aucune solution satisfaisante n'est proposée pour la prise en compte de telles structures. Nous nous appuyons sur le formalismes des logiques de descriptions pour représenter puis généraliser des structures prédicatives. La méthode repose sur une opération élémentaire de calcul de plus petit subsumant commun d'un ensemble de concepts et sur un ensemble d'heuristiques prenant en compte la spécificité des structures prédicatives.
Dans une deuxième partie, nous montrons comment appliquer notre méthode de généralisation à l'analyse de l'information. Nous détaillons le processus d'analyse de l'information, qui permet de proposer une vue synthétique de documents textuels, et qui repose sur une chaîne de traitement de ces documents. Nous montrons en particulier la nécessité d'une approche terminologique, et proposons l'utilisation de structures prédicatives comme moyen d'améliorer la finesse de l'analyse. Une expérimentation sur un corpus de résumés du domaine de l'agriculture, avec un expert documentaliste, constitue une première évaluation de notre processus de généralisation dans le cadre de l'analyse de l'information : le résultat est un accès plus efficace et plus pertinent aux informations contenues dans un corpus volumineux de textes scientifiques, grâce à une vue synthétique de ces textes.

Mots-clés: intelligence artificielle, terminologie, généralisation, structures prédicatives, logique de descriptions, analyse de l'information

Abstract

We present an original method for structuring predicates structures using generalisation, and then apply this method to analysis of informational content of scientifics texts.
In the first part, we present our method of predicate structures generalisation. A review of existing generalisation methods, which belong to the conceptual clustering domain, shows that there is no satisfying solution for such structures. We use the description logics formalism to represent and generalise predicate structures. Our method uses the operation of least common subsumer of a set of concepts, and defines heuristics specific to predicate structures.
In the second part, we apply our generalisation method to information analysis. We detail the information analysis process, which results in a synthetic view of textual documents. We show that a terminological approach is necessary, and use predicate structures as a way to improve the analysis process. We present our experimentation, with an expert, on abstracts in agriculture, which constitutes a first evaluation of our generalisation process in the framework of information analysis : the result is a more efficient access to information contained in corpus of scientific texts.

Keywords: artificial intelligence, terminology, generalisation, predicate structures, description logics, information analysis