Chapter 8
Regularity, Document Generation, and Cyc

Hafedh Mili[1] François Pachet[2]
1. Laboratory for the Acquisition and Representation of Knowledge (LARC[1])
Department of Computer Science
University of Quebec at Montreal
P.O. Box 8888, ''Downtown''
Montreal PQ H3C 3P8, Canada
E-mail: Hafedh.Mili@uqam.ca

2. LAFORIA-IBP
Université Paris 6
4, Place Jussieu
75252 Paris Cedex 05, FRANCE
E-mail: pachet@laforia.ibp.fr

Abstract

We are interested in building and maintaining semantic nets in general, and hierarchical semantic nets in particular. In [Mili, 1988], we developed a model of hierarchical semantic nets that generalizes taxonomic models by replacing the concept of property inheritance by a more general behavior of properties that we called regularity [Mili & Rada, 1990a]. We are also interested in the generation of structured documents from hypertext. We argue that authors structure their descriptive documents based on personal but systematic traversals of a model of the domain of discourse [Mili & Rada, 1990b]; when that model consists of a semantic net that exhibits regularity, the traversal can be rationalized and described concisely. Within the context of our research, we acquired an electronic copy of the Cyc knowledge base [Lenat & Guha, 1990] to: 1) use the semantic net underlying Cyc to support the generation of argumentative and explanatory documents, and 2) identify regularity patterns. Interestingly, the sheer size and richness of Cyc posed challenging performance problems to its designers, who had to constrain proof and inference procedures in such a way as to make it ill-adapted to

[1]LARC, in French, stands for Laboratoire pour l'Acquisition et la Représentation des Connaissances.

the kind of open-ended lengthy logical inferences required for generation of argumentative text. The study of regularity patterns in Cyc led us to generalize the concept of regularity, and to formulate a number of hypotheses about the semantic structure of Cyc, and of common sense knowledge in general, proving once more that regularity is a powerful tool for managing the complexity of large knowledge bases.

1. Introduction

In this chapter, we describe ongoing research at the Laboratory for the Acquisition and Representation of Knowledge of the University of Québec in Montréal. LARC boasts a dozen professors of computer science and statistics who work on different aspects of knowledge acquisition, modeling, and manipulation. Some of the representation formalisms studied include logic, semantic networks, frame languages, and conceptual graphs. Some of the applications we are considering include the semantic modeling of databases, tutorial systems, textual and software information retrieval, and multi-media databases. Over twenty graduate students gravitate in the lab at any point in time, and we regularly host visiting researchers and post-doctoral fellows.

The author(s) work on knowledge representation focuses on semantic nets, with an emphasis on hierarchical semantic nets. We are particularly interested in the representation and manipulation of manually-built hierarchical semantic nets. In previous experiments, such hierarchies proved to be useful in performing a number of "intelligent" information retrieval tasks (see e.g. [Mili & Rada, 1988]). These experiments also highlighted the need for constantly updating and maintaining such hierarchies to account for the evolution in the domain of knowledge they cover. In [Mili, 1988], we proposed a model of semantic hierarchies that generalizes taxonomic models, by replacing attribute/property inheritance by a more general behavior that we called regularity [Mili & Rada, 1990a]. This model was tested on two medical knowledge bases of 100 and 300 complex "frames", respectively, developed using concepts from the Medical Subject Headings (MeSH) thesaurus [NLM, 1986]. Subject experts confirmed that the few exceptions to regularity that we observed were due to inconsistencies in MeSH, and validated the regularity-based inferences [Mili, 1988]. Since, we have been trying to validate regularity in bigger, more complex knowledge bases that deal with areas other than biomedicine.

We are also exploring novel ways in which semantic networks may support hypertext system functionalities, with a particular interest in the generation of structured documents from hypertexts. Broadly speaking, a hypertext is a collection of text blocks connected by links, with more or less rich semantics [Rada, 1989]. A number of hypertext models have been proposed in the literature, leading to a proliferation of hypertext system architectures [Rada, 1991]. We are interested in hypertext where textual blocks are connected via independent semantic (vs. lexical) networks; depending on the model, textual blocks may point to (be

indexed by) either nodes or links [Rada, 1991]. By studying a number of descriptive documents (medical treatises), we were able to formulate and validate the hypothesis that authors structure their documents according to a personal, but systematic traversal of a model of the domain of discourse [Mili & Rada, 1990b]. In medicine, a number of such models exist as semantic networks of various kinds, including disease taxonomies, anatomical hierarchies, etc. [Mili & Rada, 1990b]. When the semantic network exhibits regularity, the traversal strategy may be explained, and described clearly and succinctly [Mili & Rada, 1990b]. We are trying to generalize this model of document structuring to argumentative documents. We hypothesize that with such documents, the underlying semantic network involves assertions and predicates, and the traversal consists of a proof procedure [Mili & Rada, 1990b].

Within the context of a collaboration with the AI lab of the Microelectronics and Computer Technology Corporation (MCC), led by Doug Lenat, we acquired an electronic copy of the Cyc knowledge base [Lenat & Guha, 1990] to support our research. Thanks to its complexity, richness and sheer volume, Cyc seemed to be the ideal platform to test our models and to explore new ones. We intended to use Cyc for two purposes: 1) using its underlying semantic/logical network to support the generation of explanatory and argumentative texts, and 2) test regularity patterns for the hierarchical relationships in the knowledge base. The sheer size and breadth of Cyc, which is one of its most attractive features, raised serious performance concerns, forcing its implementers into making a number of design optimizations that curtailed its flexibility and made our job a bit more difficult. First, a number of representation choices and inference optimizations, meant to alleviate the combinatorial explosion of inferences that Cyc could draw from a simple fact, made Cyc ill-adapted for the kind of deep logical inferences that are required for the generation or argumentative documents. Further, inference efficiency concerns led Cyc designers to adopt a hybrid representation that de-emphasizes the declarative representation style, for the benefit of a more custom-tailored, procedural style [Lenat & Guha, 1990], obscuring some of the regularity patterns which would have otherwise been more visible. However, by studying these patterns, we were able to identify a more general and powerful form of regularity, and to formulate a number of hypotheses about the epistemological structure of the knowledge base, illustrating once more the power of regularity as a means of managing the complexity of knowledge bases.

In the next section, we define regularity and describe some of the regularity-based inferences. In section 3, we discuss the problem of hypertext-based document generation, and show why we felt that Cyc would adequately support the generation of argumentative documents. We introduce Cyc in section 4. The ontological and technical choices that affect our applications are discussed in section 5. The study of regularity patterns is described in section 6. We conclude in section 7.

2. Hierarchical Semantic Nets and Regularity

Semantic networks are at the center of a number of "intelligent" information processing systems. However, the cost of building and maintaining them constitutes a true bottleneck to the development of such systems. We have long been interested in reusing existing knowledge sources which may have a poor structure, as compared to the toy knowledge bases that are typically artfully crafted in AI labs, but which could serve as a skeleton that could be enriched more or less automatically [Mili, 1988]. Such sources include thesauri or classification structures which are laboriously developed by domain experts, and used for the classification and retrieval of bibliographic documents. In previous work, we have developed a number of syntactical and structural methods for building [Mili & Rada, 1987] and maintaining [Mili & Rada, 1988] hierarchical semantic nets. "Maintenance" consists essentially in placing new concepts in the hierarchy. Experiments showed the limitations of such methods, including in deceivingly simple cases [Mili, 1988]. We have since been interested in exploring construction and maintenance methods that are based on a formal characterization of the semantics of hierarchical relationships.

The AI literature abounds with taxonomic models [Schmolze & Lipkis, 1983], [Fisher, 1987], [Lebowitz 1986, Lebowitz 1987]. We distinguish between two kinds of models, which we call *inductive* and *axiomatic*. The KL-ONE language and its descendants are typical of the axiomatic approach [Brachman, 1985]. In KL-ONE, the taxonomic relationship between concepts -- called subsumption -- is defined in terms of a predefined set of primitive relationships between concepts' properties; such a definition supports a classification algorithm that places a new concept in a KL-ONE taxonomy by comparing its properties to those of concepts in the taxonomy [Schmolze & Lipkis, 1983]. The inductive approach is illustrated by the UNIMEM system, proposed by Lebowitz [Lebowitz, 1987]. UNIMEM uses conceptual clustering methods to build classification hierarchies. The axiomatic methods have the advantage of a clear formalism, and solid cognitive foundations. However, they don't have the flexibility required to handle semantic hierarchies other than taxonomies. The inductive methods have that flexibility, but lack a clear theoretical foundation, and cognitive plausibility.

In this section, we describe a model of hierarchical relationships based on the general observation that hierarchical relationships between concepts reflect relationships, often hierarchical, between concepts' properties. This phenomenon, which we call *regularity* is a generalization of inheritance. We propose a model of hierarchies-- called DC model for Description-Context-- which is based on regularity in the same way that taxonomic models are based on inheritance. Finally, we describe a number of regularity-based inferences, and discuss some of their uses.

2.1. Example of regularity

Consider the hierarchy of eye diseases shown in Figure 1. This hierarchy is part of the *Medical Subject Headings* (MeSH) thesaurus [Mili & Rada, 1988]. MeSH is a classification structure developed and maintained by the (U.S.) National Library of Medicine, to support the operations of its MEDical Literature Analysis and Retrieval System (MEDLARS). MEDLARS can be accessed worldwide using MEDLARS ON-LINE, or MEDLINE. MeSH contains over 15,000 concepts divided among 15 categories, including Anatomical Terms and Diseases. The concepts within each category are organized in hierarchies based on the Broader-Term relationship. This relation is fairly general, and encompasses taxonomic relationships as well as other kinds of hierarchical relationships [Council, 1988]. The subhierarchy shown in Figure 1 belongs to the Diseases category.

In a medical knowledge base diseases may be described by a number of properties including their 'Location's, which are the body parts affected by the diseases, their Symptoms, their Etiology (causes), etc. The locations of the eye diseases of Figure 1 are implicit in their names. For example, the Location of Conjunctival Diseases is the Conjunctiva. If we represent disease locations in a Part-Of anatomical hierarchy, we get the hierarchy on the right hand side of Figure 1. It appears that whenever a disease A has a Broader-Term a disease B, then the Location of A is Part-Of the Location of B. We say that the Location property is regular with respect to the relation Part-Of. We have identified a number of instances of regularity both in MeSH and in other classification structures [Mili, 1988].

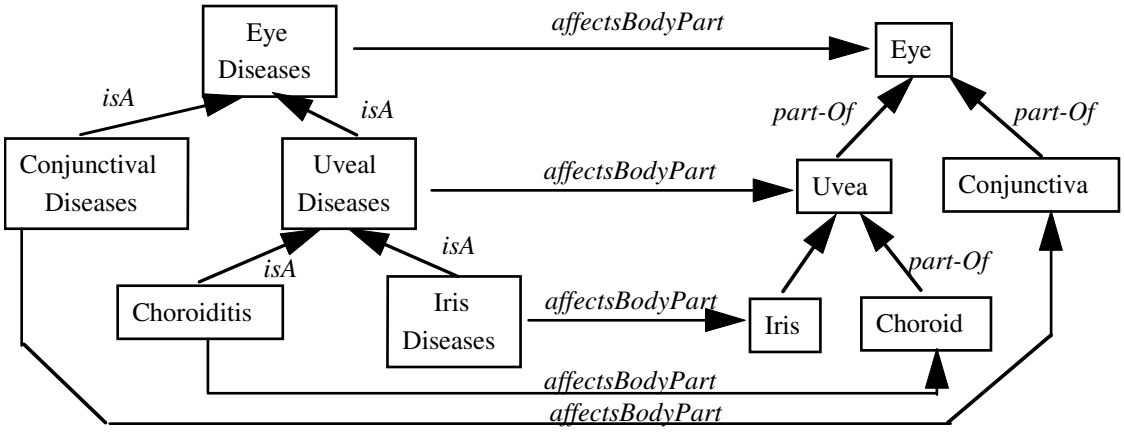

Figure 1. The subhierarchy of eye diseases and the anatomical sub hierarchy of the eye. Property "affectsBodyPart" is regular with respect to the relation "part-Of".

2.2. Definitions

Mathematically, regularity can be characterized as follows. Let N be the set of concepts (nodes) in a hierarchy, and ‘‘Lower-Than’’ a hierarchical relationship between elements of N[1]. A property (or attribute) of the concepts in N can be seen as a binary relation between the elements of N and the permissible values for that property. Let F be a property, and P the set of permissible values for F. We have F $\prod$ N X P. Let r be a binary relation defined in P. We have r $\prod$ P X P. With each relation r $\prod$ P X P, we associate a relation R $\prod$ 2^P X 2^P which is defined as:

A,B $\prod$ P and (A,B) R + $\forall$ a A, $\exists$ b B such that (a,b) r.

R is called the set relation associated with r.

Definition (1)

Let F be a multi-valued property, r $\prod$ P X P . F is regular with respect to r iff, $\forall$ *n1, n2* *Dom(F)*,

(1) n_1 Lower-Than n_2 -> ($F(n_1)$,$F(n_2)$) R

where R is the set relation associated with r. F defines a graph homomorphism between (Domain(F), Lower-Than) and *(U / {n Dom(F)} {F(n)}*,R)

Notice that when F is a function, F(n) is a singleton, and we can replace "($F(n_1)$,$F(n_2)$) R " in the implication above by "($F(n_1)$,$F(n_2)$) r". In the next section, we discuss some of the properties of regularity, and some regularity-based inferences.

2.3. Properties

First, we show that inheritance is a special case of regularity. Generally speaking, a property F is inheritable if and only if, for all n_1 and n_2, we have:

$$n_1 \text{ is-a } n_2 \varnothing F(n_1) \prod F(n_2)$$

where "is-a" is the generic name used for taxonomic relationships, and $F(n_i)$ represents the image of n_i by F. In Other words, F is inheritable iff it is regular with respect to inclusion! Note that inclusion ($\prod$) is the set relation associated with equality (=), i.e.:

---

[1]We shall use the name ‘‘Lower-Than’’ to refer to any kind of hierarchical relationship, be it ‘‘Broader-Term’’, ‘‘is-a’’, or any other relationship.

$$A \prod B \varnothing (\forall a \quad A) (\exists b \quad B) \text{ such that } a = b$$

In KL-ONE, if A "subsumes" B, then the functional roles of A must be in one of four predefined modification relationships to the corresponding functional roles of B [Brachman and Schmolze, 1985]. Thus, we have:

$$n_1 \text{ subsumes } n_2 \quad \varnothing \qquad (F(n_1),F(n_2)) \quad R_i$$

where R_i , for i=1,...,4, is one of the modification relationships.

Regularity supports a method for computing default values in a way that inheritance[1] supports a method for computing default values within taxonomies. We call such a method *expansion*. Simply put, expansion adds pairs (n,F(n)) in such a way as to preserve the regularity of F with respect to (w.r.t) R. Formally, let F be a regular property w.r.t R, and assume that we didn't know the value(s) of F(n); expansion assigns n the set S such that:

$$\text{(2.a) } (\forall g) \quad n \quad \text{Lower-Than} \quad g \quad \varnothing \quad (S , F(g)) \qquad R$$

$$\text{(2.b) } (\forall g) \quad g \quad \text{Lower-Than} \quad n \quad \varnothing \quad (F(g) , S) \qquad R$$

If such a set existed, by taking F(n) = S, we preserve the regularity of F w.r.t R. The reader can show that if n had no descendants, of if no descendant of n had known values for F, then a set S satisfies (2.a) if it is included in $\leftrightarrow$ / (n Lower-Than g) R^{-1} (F(g)) where R^{-1} (F(g)) is the set {p Œ P | $\exists$ q F(g) such that (p, q) r} [Mili, 1988]. This type of expansion is called downward expansion. Note that multiple inheritance in taxonomies may be described in terms of downward expansion w.r.t to set inclusion ($\prod$). In this case, we simply case the intersection of the "inherited" sets of values, i.e. S = $\approx$ / (n Lower-Than g) (F(g)). This kind of expansion is called upward expansion [Mili, 1988]. When there are known values for F for both ancestors and descendants of n, then a set S satisfies (2.a) and (2.b) iff:

$$\approx_{g \text{ Lower-Than } n} R(F(g)) \prod S \prod \leftrightarrow_{n \text{ Lower-Than } g} R^{-1} (F(g))$$

Naturally, such a set exists only if:

---

[1]The word "inheritance" has two meanings. Inheritance as a phenomenon, expressed by implications such as the ones shown above, and inheritance as an inference to be used, when actual values are not known.

$$\approx_{g \text{ Lower-Than } n} R(F(g)) \prod \leftrightarrow_{n \text{ Lower-Than } g} R^{-1}(F(g))$$

Bidirectional expansion assigns to n the smallest set satisfying these conditions, i.e. $\approx_{g \text{ Lower-Than } n} R(F(g))$.

Expansion is reliable because regularity relations (relation R above) are often more precise than the inclusion relationship, e.g., or the-- inevitably-- generic/permissive modification relationships used in KL-ONE. This is true in part because we do not attempt to define a predefined set of relationships that would encompass all hierarchies; instead, we use relations that are specific to the hierarchy at hand. This data-dependence has the two-fold advantage of flexibility and precision.

From a semantic point of view, expansion is justified to the extent that inheritance corresponds to a fundamental property of hierarchies. We argue that this is the case: regularity relations reflect the semantics of the underlying hierarchical relationships. In other words, regularity relations constitute the semantic primitives of the underlying hierarchical relationships, in the same way that modification relations in KL-ONE underlie the taxonomic subsumption relationship. Practically, given a hierarchy H whose concepts are described by properties $F_1,...,F_k$, which are regular w.r.t relations $R_1,...,R_k$, respectively, two fundamental questions come to mind. First, which of the regularities correspond to a semantic primitive of the underlying hierarchical relationship? Indeed, there are cases where the regularity of a property is fortuitous, and non-essential. Consider the command structure of an organization, where we use the "reports-to" relationship between the different positions. It will often be the case that the age of an officer/executive be lower than that of his superior. In this case, we can say that the Age property is regular with respect to the relation (total order in this case) Smaller-Than (<). However, this is a simple consequence of the fact that administrative positions are assigned based on a number of competencies that tend to grow with age (e.g. experience). However, hiring decisions are not based on age. Second, assume that we have identified the "real" semantic primitives; what is the exact logical expression of the underlying hierarchical relationship in terms of regularities of properties? We developed a model of hierarchies-- which we called Description-Context (DC) hierarchies-- which addresses this concern [Mili, 1988]. A context describes a point of view or perspective on the set of concepts included in the hierarchy.

For the purposes of this chapter, we will be content to mention that the model supports a classification algorithm that preserves contexts [Mili, 1988]. We will see in section 8.3.2 an example that illustrates the precision of the DC model, as compared to the more traditional taxonomic models.

3. Generating documents from hypertext

3.1. Hypertext

The idea of hypertext is commonly attributed to Vannevar Bush. In his seminal paper [Bush 1945], he proposed a mechanical device that physically arranged a set of documents in such a way that they could be presented in different sequences, reflecting various semantic relationships between them; such an organization was to mimic the organization of long-term semantic memory. Except for the fact that today's hypertext systems are electronic, the initial design and its rationalization have pretty much lived on unchanged. Simply speaking, a hypertext is a set of textual blocks connected by explicit links having more or less rich semantics [Rada, 1991]. A hypertext system is a software package supporting the presentation, navigation, and otherwise manipulation of hypertexts. There are a number of hypertext models in the literature, and as many corresponding hypertext system architectures [Rada, 1991]. We are interested in hypertext models where the connections between text blocks are done via an independent semantic (versus lexical) network; depending on the models, textual blocks are connected to either nodes or to links of the semantic network [Rada, 1991].

A hypertext system enables its users to view a hypertext through a number of alternative navigation paths, as opposed to the only available sequential navigation imposed by the immutable structure of hardcopies or sequential electronic files. However, this flexibility has always been a mixed blessing: users have the tendency to quickly lose track of where they are and where they are going. According to Van-Dam, one of the co-creators of the Hypertext Editing System-- one of the first hypertext systems:

> «We already started getting the notion that the richer the hypertext, the greater the navigational problem. But we arranged careful demos in which we knew exactly where we had to go, and people were impressed...» [Van Dam, 1988]

We argue that despite its attractiveness, the presumed cognitive plausibility of navigation is misguided for the case of hypertext. For instance, even if we accept the semantic network model of semantic memory [Quillian, 1968] and the-- navigational-- spreading activation model of cognitive associations/comparisons [Collins & Loftus, 1975], the "spreading" is supposed to be parallel, subconscious, and self-regulating in the sense that the "activation" (attention or focus) fades progressively as we move away from the original concepts to be compared [Collins & Loftus, 1975]. Yet, in a hypertext system, navigation is sequential, conscious/attentive, and nothing tells the user that he has strayed away or calls him/her into line.

Accordingly, a number of researchers tried to complement the atomic navigation functionalities offered by hypertext systems with more structured navigation functionalities

that enable users to choose a navigation strategy and to pursue it during their explorations [Rada, 1991]. These functionalities do not affect the flexibility of the underlying navigation mechanisms, to the extent that: 1) users are free to choose the navigation strategy that suits their needs, and 2) the two kinds of navigation may co-habit in the same system, enabling users to "digress" if they wish. To identify the strategies that we wanted to offer to users, we studied the structuring strategies that authors/writers themselves use to organize their prose. These strategies are discussed in the next section.

3.2. A model of documents

Non-fiction writers appeal to a number of models of their areas of discourse. Such models are reflected in the structure of their books. We tried to study such structures to systematize their generation. Inherent to the outline of a book are two kinds of relationships: 1) hierarchical relationships of textual containment between a section of text and its subsections, and 2) precedence relationships between sibling subsections, as reflected by their sequential arrangement. We studied the structure of several medical documents to characterize the two kinds of relationships [Mili & Rada, 1990b]. Our choice of the medical literature was motivated by two factors: 1) the medical expertise of one of the authors[1], enabling us to understand and interpret some of the structuring choices made by authors, and 2) the maturity of medicine as a scientific discipline, and the stability of its models.

The precedence relationships usually embody one of three ordering relationships: 1) incidental or non-essential ordering, 2) temporal relationships, and 3) logical relationships. Temporal relationships, prevalent in descriptive documents, are used when the subject of discourse of one section precedes (or succeeds), time-wise, the subject of the next section. This is illustrated by the organization of the family medicine classic Clinical Obstetrics, whose chapters are organized as follows:

1. **Early development of the fertilized egg**
2. **Physiological relations of the mother**
3. **Antepartum care**
4. **Labor**
5. **Puerperal management of the mother**
6. **Care for the neonate**

[1]Roy Rada was a Medical Doctor before earning a Ph.D. in Computer Science

Logical relationships are found in argumentative documents, and are discussed in the next section.

Textual containment relationships generally depend on the precedence relations among the subsections. In particular, when the ordering of the subsections is incidental, the relationship between a section and its subsections is binary. Otherwise, we argue that the relationship is (n+1)-ary, where n equals the number of subsections. Binary relationships are consistent with an incidental ordering of the subsections. In Harrison's Internal Medicine we identified binary, hierarchical relations and binary, frame/slot relations. Binary, hierarchical relationships are usually based on a specific model of the topics covered. For example, a taxonomy of diseases may be based on a taxonomy of organ systems, as in Harrison's:

8. Diseases of the organ systems

8.1. Diseases of the respiratory system

8.1.1. Diseases of the upper respiratory tract

8.1.2. Diseases of the Pleura, Mediastinum, and Diaphragm

8.2. Diseases of the hepatobiliary system

It is interesting to note that the relation between "respiratory system" (or "hepatobiliary system") and "organ system" is an "is-a" relationship, whereas the relationship between "upper respiratory tract" (or "pleura, mediastinum, and diaphragm") and "respiratory system" is a "part-of" relationship. In other words, the "location" property (see section 8.2.1 of this chapter) of sections, is regular with respect to "is-a" for the first level (8.1.∅ 8. and 8.2. ∅ 8.), and with respect to "part-of" for the second level (8.1.1. ∅ 8.1. and 8.1.2. ∅ 8.1.). In a typical taxonomic model, the hierarchical relationship between sections is invariably "is-a", independently of the level. In our DC model of hierarchies, because of the differences in regularity relations, we are dealing with two distinct hierarchical relationships. To see the difference, consider the following outline where we added a third subsection to 8.1. (8.1.3.). We believe that the relationship between 8.1.3. and 8.1. is perceptibly different[1] from that between 8.1.2. (or 8.1.1.) and 8.1.

8. Diseases of the organ systems

---

[1]In fact, our DC model of hierarchies can accommodate cases where the hierarchical relationships within a hierarchy are "slightly" different; acceptables differences correspond to the case where the relationships within one level of the hierarchy are a specialization of the relationships at the upper levels [Mili, 1988]. In this example, that would correspond to saying that "part-of" (e.g. between "upper respiratory tract" and "respiratory system") is a specialization of "is-a" (e.g. between "respiratory system" and "organ system"), hence the distinct impression that "Diseases of the fetal respiratory system" is out of place.

8.1. Diseases of the respiratory system

8.1.1. Diseases of the upper respiratory tract

8.1.2. Diseases of the Pleura, Mediastinum, and Diaphragm

8.1.3. Diseases of the fetal respiratory system

8.2. Diseases of the hepatobiliary system

Other kinds of binary relationships may correspond to that between (the description of) a concept and (the description of) its properties, as illustrated by the following example, excerpted from the same chapter:

8.1.1.2. Diffuse infiltrative diseases of the lung

8.1.1.2.1. Pathogenesis

8.1.1.2.2. Clinical manifestations

8.1.1.2.3. Diagnosis

8.1.1.2.4. Treatment

8.1.1.2.5. Prognosis

For the case of binary relationships between a section and its subsections, the textual precedence between sibling subsections is either arbitrary or based on subjective factors. For example, when the relationship between a section and its subsection is a specialization, the ordering between sibling subsections may be purely alphabetical. When the relationship between a section and its subsections is n-ary, the textual precedence relationship between sibling subsections is usually inherent in that n-ary relation. Take the example of a history book about North Africa:

3. (The history of) North Africa

3.1. The pre-islamic era

3.2. The islamic conquests

3.3. The French colonization

3.4. The independence movements

The relationship between a section (e.g. 3.2.) and the chapter (3.) is that of temporal inclusion, and this relationship can be perceived as binary. However, if we impose the constraint that we have to cover the entire period implicit in the chapter, then the relationship

becomes quinary. We observed a similar pattern when textual precedence is logical: a chapter may represent a thesis to be proven, and different subsections may elaborate different steps of the proof. In this case, the table of contents resembles a depth-first traversal of and AND/OR tree.

The previous examples showed that the structure of documents reflects, to varying degrees, the structure of the domain of discourse. The sequential nature of traditional media requires some sort of linearization of the structure of the domain. As shown with the "Diseases of the organ system" examples, the more systematic the linearization, the more coherent the presentation. We argue that a coherent presentation facilitates assimilation because it creates expectations in the user's mind, facilitating the integration of the new information acquired during reading. Finally, as shown with the various sections of the Harrison's Internal Medicine book, authors may choose to traverse different relationships of the domain model at different levels of the structure of the document. For example, the structure of chapter 8 of the book may be described by the following three rules:

- The textual containment relationship of the first level is based of the regularity of the "location" property with respect to the "is-a" relationship; the textual precedence relationship is alphabetical,
- The textual containment relationship of the second level is based on the regularity of the "location" property with respect to the "part-of" relationship; the textual precedence relationship is alphabetical,
- The textual containment relationship of the third level relates a concept to its properties ("slots"); where/when applicable, use temporal ordering between the subsections.

We could imagine a representation of Harrison's Internal Medicine where only the subsections of the lowest level are represented by actual textual blocks corresponding to the (description of the) properties of the concepts-- in this case, diseases. Given a navigation/outline specification/description language that enables us to express the above rules, we could navigate the book within a hypertext system using this outline description, or, alternatively, automatically generate a hardcopy to consult off-line. This kind of navigation can co-exist with other kinds of navigation strategies that follow cross-references, bibliographic references, etc.

3.3. The generation of argumentative documents

We use the term "argumentative documents" to denote documents that develop a thesis, or prove an assertion, as opposed to descriptive documents such as the medical treatises studied in the previous section. The argumentative versus descriptive nature of documents is a global property to the extent that an argumentative document could (and must!) contain descriptive elements, and vice versa. To revisit an example from the previous section, one would expect

to find some justifications in the description of the “Pathogenesis” or “Diagnosis” (i.e. in the corresponding document sections) of diseases, especially considering that the books target an audience of specialists. We argue that the archetypal structure of an argumentative document consist of a hierarchization, followed by a linearization of a proof procedure for the main thesis of the document. We first illustrate this hypothesis by an example, and then discuss the problems that may occur in practice, and the reasons that led us to believe that the Cyc knowledge base could address those problems.

Consider the following logical expressions, and assume that they are all true:

$A \wedge B \varnothing C$, $D \varnothing E$, $E \varnothing A$, $F \vee G \varnothing B$, F, D

Assume now that we have six blocks of text, each explaining or stating one of the six expressions; we could say that each bloc is indexed or catalogued with the corresponding logical expression. Suppose now that we had to “state” or, rather “prove” why C is true. A possible document structure might be:

«C» is true

1. «A» is true

1.1. «D» is true

1.2. «D ∅ E» is true

1.3. «E ∅ A» is true

2. «B» is true

2.1. «F» is true

2.2. «F V G ∅ B» is true

3. «A ∧ B ∅ C» is true

Such a structure may be generated from the proof tree of C, using some additional structuring conventions. For example, regarding the textual precedence of sections, we choose to proceed from antecedents/hypotheses to consequences/conclusions; we could have chosen the opposite. As for the textual containment relationships, the general rule seems to say that the subsections of a given section S represent: 1) the rule(s) having S as a consequence, and 2) the antecedents of such rule(s). We made an exception in the example above for the case of section 1 («A») in order to not have a narrow and deep structure; this could be expressed by a condition on the number of antecedents of rules: if the number of antecedents of a given rule is below a given threshold, we "flatten" the proof tree. Without this exception rule, section 1. would look as follows:

1. «A» is true

1.1. «E» is true

1.1.1. «D» is true

1.1.2. «D ∅ E» is true

1.2. «E ∅ A» is true

This example shows that aesthetic and ergonomic considerations can come into play during the generation of argumentative documents, and can be easily accommodated using simple rules. We could even imagine adding an introductory subsection to each section, giving an overview of the "proof" contained in the section, etc. However, the real challenges reside somewhere else. First, we have to have a substantial logical knowledge base. An argumentative document generator in a specialized domain is not much different from an expert system with an explanation facility, and as such, suffers from the same problems as expert systems with regards to explanations [Clancey, 1983]. Among other things, we have the problem of the lengths of inferences. For the case of INTERNIST, the derivation of a diagnosis may involve the firing of hundreds of rules. This large number is due, in part, to the fact that several rules are utterly "uninteresting", but are needed to connect antecedents to conclusions [Clancey, 1983]. Note also that the explanations generated by expert systems are, in fact, justifications, and are of very little value to novices. It is interesting to note that when INTERNIST was changed to be used for teaching purposes, it was augmented with a hypertext system component that explains the rationale behind some of the rules [First et al., 1985].

The case of non-specialized domains is even more challenging. One of the authors observed a journalism student assemble interview notes and quotes for a socio-political study[1]. The student had interviewed some forty clerics and lay people of various social stature, and collected some their most poignant/revealing quotes. When the time came to write down the report, she realized that in addition to reconciling the divergent views she collected on the topic, she also had to find the right quotes to illustrate the positions she was conveying. Lest we oversimplify, if we had indexed each quote with an assertion, and if we had a rule base and a theorem prover, it may have been possible to submit the main thesis of the report as a statement to be proven by the theorem prover, and let it collect the quotes along the way.

---

[1]The study was concerned with the role of the church in the social and political makeup of modern Quebec.

Notwithstanding the problems mentioned above (length of proof procedure), and linguistic/stylistic problems (paraphrasing rules, and connecting proof stages), we still have two quasi-intractable problems. First, unlike argumentative systems in specialized domains where the domain of discourse is rather narrow, and where the rules are, all things considered, rather synthetic, the system we need in this case has to have some common sense. While a specialized system can get by with a relatively small number of rather coarse rules, a general-purpose system needs a large number of finely grained rules. The second problem is formal, and is related to the monotonic nature (or lack thereof) of the underlying logical system. First, several of the quotes contradicted each other[1], reflecting diverging opinions; a regular (monotonic) "theorem prover" would fail. Further, common sense knowledge is ridden with defaults, and is essentially non-monotonic in nature.

It seemed to us that the Cyc knowledge base would address the above problems: 1) it embodies common sense knowledge, and contains relatively fine-grained knowledge, and 2) it uses several default inferences that seek the most "plausible" inferences [Lenat et al., 1990]. For these reasons, we felt that Cyc would enable us to explore a number of semantic issues involved in argumentative writing. Some of these issues are illustrated using the journalistic application mentioned above. For example, objectivity (impartiality, or "fair reporting") would require us to account for the various opinions. If in the process of proving thesis T, we also prove its opposite ÿ T, but that the proof for T is more plausible than the proof of T, we could still consider T to be the main thesis of the document, but nuance it with the arguments against it. Another interesting case which humans (and specially reporters and politicians) seem to handle quite well, has to do with incomplete proofs, where we overlook some rule antecedents[2], but nuance the conclusions.

Unfortunately, as we see in sections 8.4 and 8.5, Cyc turned out to not be as well adapted to the kind of uses we had in mind, as we first thought, and the issues raised above remain unexplored for the time being. The difficulties are not epistemological in nature; they are related to a number of constraining implementation choices that had to be made by its designers because of its sheer size and complexity.

4. Cyc

The Cyc system, developed by the team of D. Lenat, is the most ambitious attempt to date at representing a substantial amount of common sense knowledge. This project began in 1984, and was initially supposed to span over a period of ten years. Available documents on the Cyc system and the Cyc knowledge base include a book written at mid-term [Lenat & Guha,

[1]Including from the same person... :-)

[2]Scientists turn unproven antecedents into hypotheses

1990] e the main representation mechanisms and desig ontology. These documents may not be consi stem which is still today (1995) in its devel er precise indication of the philosophical state and feel of the system. The Cyc effort has been by the AI community. Strangely enough, the cr phical assumptions of the team of Douglas Lenat pects of the undertaking.

Indee , and provoking [Lenat & Feigenbaum, 1991] . These assumptions are twofold. First there are assumptions on the *nature* of intelligence. The main one being that intelligence requires a large quantity of knowledge to manifest itself. According to Lenat, the main consequence is unfortunate, but definitive: there cannot be any "Maxwell equation of thought". The only way to exhibit artificial intelligence is to build a large quantity of knowledge that will serve as a support for intelligence.

On the other hand, there are hypothesis on the nature of this prolific knowledge that can serve as a support for primitive intelligence. The main argument is that the acknowledged brittleness of first generation expert systems relies on the inconsistency - and lack of semantics - of the predicate and constants manipulated by these systems. Lenat identifies a "common sense" layer of knowledge that could fill the gap between all expert systems, by providing some sort of "semantic glue". This layer includes knowledge of a particular kind, that is never explicitly written in textbooks or dictionaries, and therefore which is hard to define. Common sense is the opposite of "expert knowledge". For example it includes what human needs to know in order to read and understand dictionaries or encyclopedias. Typical common sense knowledge chunks are: " children are younger than their parents" ; "owning something implies owning all its parts" ; "water flows downstream" ; "Birds lay eggs" ; "clothes hide bodies", and so forth[1]. This layer of knowledge is supposed to be logically globally consistent, i.e. common sense reasoning is assumed to be representable within a logical framework. Furthermore, Lenat does not attempt at representing exhaustively all common sense knowledge, but assumes that there is a "critical" mass of knowledge that can be manually (chirurgically) produced, from which it will be possible to build automatic text comprehension systems that will perform the rest of the work by themselves.

These various hypothesis have been widely criticized [Smith, 1991]. The technical aspects have received less attention, mainly through a late critics of the book itself seen as a scientific document [CycBookReviews, 1993], including answers to the critics by the authors of the

[1] Of course, all the difficulty is to explicit this "so forth".

Cyc system, in which a systematic engineer standpoint if taken. On a different tone, the interesting book [Conversations, 1994] (chapter 4) is a transcription of a workshop gathering Lenat and major AI researchers (McDermott, Steels, Chandrasekaran, Clancey, Mitchell, Cohen). Here again, even if the arguments on both sides are impartial and fascinating (nature of knowledge, definition of intelligence, validity of use-neutral representations), technical questions are simply ignored.

The initial descriptions of the Cyc systems [Lenat & Guha, 1990] were largely based on a "frame-oriented" view of knowledge representation. More recently descriptions indicate a shift in the representation paradigm, at least at the interface level (the so-called *epistemological* level) : talking and listening to Cyc, is mainly performed by using logical expressions, in a reified first-order logic, i.e. by writing expressions using predicates, constants, and logical variables, as well as a wealth of logical connectors for maximum expressiveness (negation, quantifications, modal operators, and so forth) [Guha & Lenat, 1994]. For various reasons the measurement of the Cyc knowledge base is not an easy task. However, it is estimated to contain a couple millions assertions (such as "when someone owns something, he/she also owns all the parts of the owned object"), talking about around 5000 constants (ranging from JamesJoyce and France to WorldWarII), 8000 collections (from the collections of all Hyundai cars to the collection of all events) and 5000 predicates (from the performsProcessType and partOf relationships to age and weight).

4.1. Classification of inference patterns

The original textbook on Cyc [Lenat & Guha, 1990], proposes a distinction between several inference patterns, based on the experience of the team in common sense knowledge acquisition. This classification aims at detecting syntactic regularities in rules, that can be used to optimize their representation, and speed up the associated inferences, as well as the bookkeeping related to truth maintenance. Around 30 classical patterns are identified, ranging from simple ones (slot inheritance, inverse) to more sophisticated ones such as the transfersThrough relation. As an example, this relation represents the following inference rule pattern:

IF (x R y) AND (y R' z) AND (transfersThrough R' R) THEN (x R z).

For instance, "owns" transfersThrough "partOf" signifies that the owns relationship "propagates" to the partsOf the object owned (e.g. owning a car implies owning all its parts, the wheels, the engine etc.). Lenat argues that most common sense knowledge may be expressed using one of these inference patterns. In rare cases, the most general inference rule is still available, but it is to be avoided as much as possible since it inherently produces the worst complexity. Since recent descriptions of the system emphasize and advocate a purely logical point of view, the distinction between inference patterns has disappeared, being

ever, we think that this classification
e effort.

have not yet been fully developed. [Lenat & Guha, 1991] proposed a could be particularly useful. These applications help people select of type of new car to buy), infer their hobbies, interests and decide which nents to convince them to buy the product), data base fields to Cyc predicates and use common sistencies, and resolve contradictions), smart point out unusual values), corporate knowledge chine translation for technical documents, and

f the Cyc research which is directly applicable to the work on context formalization [Guha, 1991]. In this theory, the knowledge base is organized into various *micro-theories* which are locally consistent, although not necessarily consistent one with each other. A number of mechanisms are introduced to compose and inherit micro-theories (an implementation of the so-called *lifting rules*). An interesting application of these mechanisms is to be found in [Pinto et al., 1995] for the representation of geographical reasoning at various levels of detail and from different standpoints (Cartesian, spherical, terrestrial).

By its sheer size and complexity Cyc is unquestionably an ideal test-bed for experimenting with our regularity methods. Before proceeding with the application of regularity to Cyc, we will provide some insights on the Cyc system from a user point of view.

5. A user's view of Cyc

We describe here our position regarding the Cyc system from a - rather naive - user's point of view. Regardless of the purely technical aspects inherent to a system of that size (bugs, interface, documentation), we identified three major problems that we feel have escaped the attention of the designers. We will now describe each of them before showing how the notion of regularity may be extended and applied in the context of Cyc to provide some answers to them.

ing the conception of the Cyc knowledge
l by an explicit constant in the system. For
of all French cars, another, KoreanCars the
r the system knows about is represented by
ennedyWasMurdered, etc.) Of course, this
proliferation of constants that could be
umber of such constants, however,
t & Guha,
1990] propose some elementary design principles. For instance, only introduce constants representing something "on which some interesting, consensual, properties have to be expressed". A typical counter-example is the predicate "dogTypeSlot", that could represent the list of all predicate that apply specifically to dogs. This list could include for instance "dogShowsWon", a typically canine slot. However, the constant is not introduced because no consensual property seems to exist for that slot, that would significantly different from other collections of slots. In ambiguous cases, Lenat provide some clues to decide for or against the reification of concepts.

Although fairly natural - especially considering the logical and frame-based background of Cyc - this position can raise serious problems, specially during the construction phase of Cyc. A strong criticism of Elkan & Greiner [CycBookReviews, 1993], p. 45, was to argue against this widespread reification, by comparing with other alternatives, using functional or compound expressions. For instance, the major of Austin could be represented either as a constant (TheMajorOfAustin) or as an expression such as (Major (Capital (Texas)). If the Cyc system is claimed to support such compound expressions (see the answers to the critics [CycBookReviews, 1993] p. 158), the issue is not only syntactical as the authors seem to believe. The standard way of entering knowledge in Cyc, and of using Cyc is to manipulate constants, which are explicitly entered by the knowledge enterers. Any decision not to reify a given concept is therefore irrevocable; the system cannot reify by itself, except in particular cases, such as the skolemization of existential variables. Regularity can provide some answer to this problem, especially with the expansion principle (section 8.2.3).

5.2. Navigating in Cyc

As in most large knowledge bases, the Cyc system induces a navigation problem, due to the large number of entities it manipulates. This problem is threefold in Cyc: the formulation of questions, the control of inference and the separation of semantically close predicates.

5.2.1. Formulation of questions

Since the appearance of the pioneer system INTERNIST [First et al., 1995], one of the first large scale expert systems, the formulation of questions has been identified as a major bottleneck of large knowledge bases. In order to be fully understood, the user of a large knowledge base must use precisely the terms manipulated by the system, and cannot expect any flexibility from it. For the user, this amounts to knowing precisely the details of the structures manipulated by the system. This problem is not far-fetched, and arises as soon as non trivial questions are asked. For instance, let us suppose the system knows about John and Paul, two instances of HumanPerson, and also knows about a car instance. Further, suppose the user wants to assert somehow that John *owns* the car. After browsing through the knowledge base (using syntactical tools), several candidates will be selected to represent the "own" relationship, say: "owns", "possess", "isLegalOwnerOf", "buyerOf", "actorIn", and so on. The problem is then to choose among these predicates the one which is most adapted to represent the desired assertion. Note that from the Cyc point of view, according to the minimality principle seen above, these slots are different precisely because they do represent different conceptual relationships : each of these slot carries different semantic information, and is used in different sets of assertions. In a way, the richness of the Cyc system comes from its capacity to distinguish between several important variations on the "own" relationship. Ideally, the system should be able to disambiguate the word "own" and find the most appropriate slot to represent this relation in the context of the expression being asserted. However, this disambiguation is not implemented for the moment. The only solution consists in browsing through the knowledge base to find - in a backward chaining fashion - all the fireable rues that each slot is likely to trigger. In our example, if we want the system to draw inferences about some possible Selling or Buying Event, then the slot owns will prove more adapted, because there exist inferences that link the predicate owns with the Selling and Buying Event associated to the financial transaction. In this case, a typical linking inference is the rule that says that "buying" something implies "owning" it, which is expressed with a reference to the slot owns, and *not* possess or isLegalOwnerOf. On the contrary, the slot possess will not be linked to these concepts, or indirectly, through a chain of inferences which is beyond the capacity of the system. The paradox is that finally, to find the right formulation of our assertion, the user will have to perform - backwards - the inferences that he wants the system to eventually draw in the first place.

5.2.2. Control of Inference

The navigation problem is further complicated by the problem of inference control. In Cyc, some rules have the capacity of generating an infinite number of dangerously prolific objects. For instance a rule states that all human beings have parents. Another one says that human beings have anatomical parts; a third one that they can breath, sleep and eat. Without some form of control, the system, given one single instance of HumanBeing could wander endlessly in the creation and contemplation of the parents of this person, their anatomical

parts, the parents of the parents, and so on. In order to limit the combinatorial explosion associated with these rules, Cyc associates an *access level* to the inference, and triggers only rules whose access-level is less than or equal to the access-level of the original request. The user is responsible for giving the appropriate access-level for its questions. Other control parameters include the number of desired answers, as well as the maximum time to spend looking for answers. In a number of cases, the choice of the right set of control parameters necessitates, once again, to anticipate the inference paths that the user expects the system to follow.

5.2.3. The semantics of slots

Like in all frame-based systems, the semantics of concepts is based on the notion of slot. Slots carry the semantics burden by linking concepts, which have a purely passive role. All the relations between concepts are represented by slots, whose values are collections of concepts. Several "meta" information are associated to slots such as their arity, the types of their arguments, and various kinds of constraints. However, it is clear that slots do not carry semantics in themselves, and make sense only within a knowledge base. Since this notion of semantics is intrinsically defined recursively, it is all the more difficult to understand it for a user. In our navigation perspective, this semantics could be represented by the tree of all possible inferences that could be drawn from this slot, combined with the rest of the knowledge base. Since this tree is in most cases infinite, it cannot be used to help user understand their semantics. However, we made the following remark: in most cases, what is needed is not so much the semantic of one particular slot, but the difference between the semantics of two closely related slots (such as "owns" and "possess"). In this case, the computation of the *differential tree* of inference is in some cases possible, and can be used for explanation purposes. These works are still in progress and consists in studying appropriate tools for constructing and visualizing differential trees.

6. Regularity in Cyc

The notion of regularity as we introduced it in section 2 appeared particularly well adapted to the navigation problem in Cyc. This is due to the fact that Cyc explicitly manipulates a large numbers of concepts deeply organized in various kinds of distinct hierarchies. The objective of Cyc is to "carve up" the world by detecting and representing explicitly these regularities. A large number of regularities in the common sense perception of the world are already represented in the Cyc knowledge base. However, these regularities are not represented as such, but are rather represented as collections of rules or constraints or constants disseminated in the system. We will now show how the notion of regularity may be applied to the Cyc knowledge base to explicitly represent these information in a more abstract way, thereby providing some answers to the problems discussed above, and provide elements for a productive use of Cyc in the context of argumentative document generation.

We will now introduce an extension of the notion of regularity as defined in section 1. This extended regularity will allow us to revisit the Cyc knowledge base from a different and enlightening perspective. We will show that some rules in Cyc tend to entertain some form of regularity (in our sense of the word), while other rules tend to break regularities.

6.1. Extended Regularity

In this section, we introduce an extension of regularity that allows to describe complex relationships between arbitrary hierarchies, expressed as arbitrary paths of inference, instead of the atomic relation of the original definition.

As an example, let us consider the concept Automobile, representing the collection of all automobiles Cyc knows about, and its various specializations: FrenchCar (itself divided into PeugeotCar, Peugeot403Car RenaultCar, and so forth): SouthKoreanCar (specialized into HyundaiCar). The Cyc knowledge base contains the slot "countryOfManufacturer" that associates any car with the country in which it was manufactured. The country is represented by a constant in the knowledge base. For instance, all instances of FrenchCar have a slot countryOfManufacturer whose value is the constant France. Similarly (or rather, regularly), all instances of SouthKoreanCar, have SouthKorea as a value for this slot.

This example clearly shows that there is a regularity between the two hierarchies: Automobile and the specialization hierarchy on the one hand, and Country and the subRegion relation on the other. This regularity is expressed by the slot countryOfManufacturer. However, this regularity is not direct, because of the relation "instanceOf" that exists between a given instance of Car and its type. In other words, the regularity we are talking about is really between the *composition* of the relations instanceOf with relation countryOfManufacturer (Cf. Figure 2).

In order to generalize regularity, we substitute binary relations of Definition (1) by arbitrary paths of inference called *access-paths*. An access path is a composition of elementary relations (slots in Cyc) that constitutes an indirect link between a source hierarchy and a target hierarchy.

Definition (2)

Extended regularity is defined as follows. Let :

- A be a collection of concepts, e.g. all the specialization of the automobile collection : FrenchCar, SouthKoreanCar, EuropeanCar, etc.

- r be a relation between concepts of A, e.g. subBrandOf or specializes

- B be a collection of concepts., e.g. all the countries (World, Europe, Asia, France, SouthKorea)

- r' be a relation between concepts of B, e.g. geographicalSubRegionOf

- r be an access-path that constitutes a link between elements of A and elements of B, e.g. allInstances°countryOfManufacturer

- B be a collection of concepts (e.g. all the specialization of the automobile collection : FrenchCar, SouthKoreanCar, EuropeanCar, etc.).

c is said to be *regular with respect to A, B, r and r'* iff:

For all n1, n2 A:

$$n1\ r\ n2 \rightarrow (c(n1), c(n2)) \quad R'$$

Where R' is the set relation associated to r'.

The preceding example constitutes a typical example of extended regularity. Note that this definition subsumes the preceding one since standard regularity correspond to access-paths reduced to elementary relations.

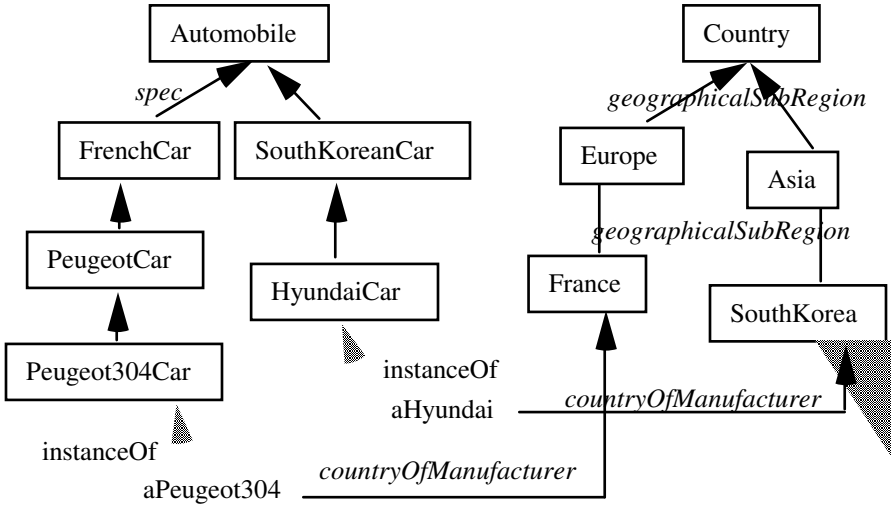

Figure 2: Extended Regularity between types of car and countries.

An other interesting example of extended regularity in Cyc is to be found between the same hierarchies and by substituting the preceding access-path

(allInstanceOf°countryOfManufacturer) by the more complex: allInstancesOf ° madeBy ° countryOfMainActivity, where:

- madeBy is a slot that links a product to its ManufacturingOrganization (an instance of AutomobileManufacturer) and

- countryOfMainActivity is a slot that links a Manufacturer to its country of activity.

This more complex regularity can also be seen as a composition of two simpler regularities: the one illustrated above and a regularity between Manufacturers and Countries expressed by the slot countryOfActivity. Indeed, FrenchCars are madeBy (instances of) Manufacturers whose mainCountryOfActivity is France. Idem for SouthKoreanCars and SouthKorea.

Theses examples show how we can "talk" about a couple of hierarchies linked in a non trivial way using an abstract property, extended regularity. Note that in the current state of the Cyc knowledge base the regularity may "work" but is not represented as such. It "works" in the sense that the appropriate inference may have been entered in the knowledge base. But these inference are represented on an individual basis. For instance, there is a rule that says that "All French Cars are manufactured in France". This rule is represented as a slot inheritance in the collection FrenchCar. A similar rule is represented in the collection SouthKoreanCar, and so forth.

The advantage of characterizing explicitly the "regularity" of these properties (or these access-paths) over representing it as a collection of special rules disseminated in various collections is obvious. We will see that this characterization may also be used to infer missing links and to detect errors.

6.2. Regularity of specialized inference patterns

We initially showed that regularity was an extension of inheritance in hierarchical networks. Similarly, we can show that some specialized inference patterns of Cyc have a corresponding definition in terms of regularity. We will take two examples: generalized slot inheritance and temporal inferences.

Let us consider the famous rule that states that all birds fly. In Cyc, this rule will be expressed (at the lower heuristic level) as a slot inheritance: any instanceOf Bird inherits the value Flying-Locomotion for its slot performsProcessType. This rule can be seen as a degenerated form of regularity since all instances of Bird will inherit the same value. However, we can rewrite the rule as a regularity pattern as follows:

A : all specialization of Bird,

r : specialization relation,

B : all the specialization of Flying-Locomotion (may be reduced the Flying-Locomotion),

r' specialization relation,

access-path: allInstances°performsProcessType.

We do not take into account the problems caused by exceptions (penguins do not fly): these are represented by explicit default rules.

Representing this rule as a regularity pattern may prove very interesting in the case of missing concepts. Since Cyc is build manually, chances of missing concepts are important. Also, in virtue of the minimality principle, only the "interesting" concepts are entered, i.e. the concepts on which specific assertions have been found and represented. This may explain some strange holes of the system. For instance, the constant EuropeanCar had no link towards the constant Europe when we made our experiments (though it may have been added later on). The expansion mechanism can be used to automatically infer the link between EuropeanCar and Europe, if the relation had been explicitly entered as regular. As we saw, however, such an inference may performed only if slot values are limited both "upward" and "downwards". This is the case if some critical mass of information has already been entered, which is consistent with the "critical mass" hypothesis of Lenat.

6.3. Regularity of relations between temporal sub-abstractions

The notion of *temporal sub-abstraction* is introduced in Cyc to represent the variations between different temporal "slices" of a given concept during its evolution. For instance, LieutenantColombo *per se* is represented by a single constant. However, different sub-abstractions will be created to represent him during various intervals of time deemed significant in a problem solving context: ColomboDuring1stEpisode, ColomboDriving, etc. This temporal sub abstraction relation is hierarchical: any sub-abstraction can in turn be decomposed into sub-sub-abstractions and so forth. This relation exhibits some regularity when it is coupled with other hierarchies of the knowledge base. For instance the fact that Colombo owns a Peugeot304 car is represented by the fact that each temporal sub-abstraction of Colombo owns the corresponding temporal sub-abstraction of the car: ColomboDuring1stEpisode owns TheCarDuring1stEpisode and so on (Cf. Figure 3). This regularity is natural and follows a co-extentionality principle which is represented in Cyc by a number of axioms, that create sub-abstractions on the fly, when needed.

Once again, there is an interpretation of this pattern with regularity. In our case, the regularity is between:

Hierarchy A : allInstances of Person, relation r : sub-abstraction,

Hierarchy B : allInstances of TangibleThing, relation r' : sub-abstraction,

Access-path : owns.

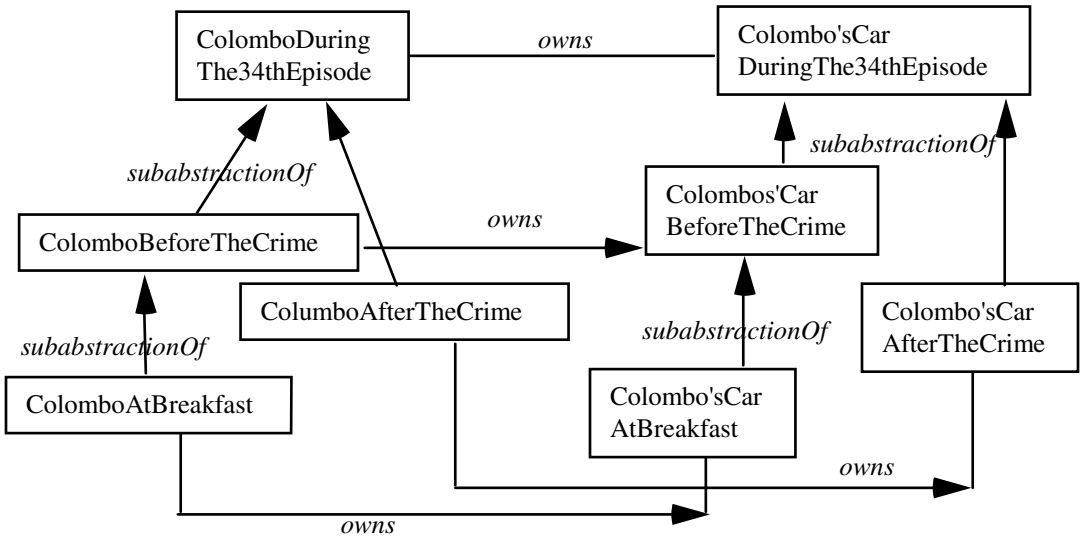

Figure 3 : Temporal sub-abstractions as regularity patterns.

Temporal sub-abstraction exhibit regularity with other relations as well, such as parents, livesIn, etc. It is not regular with all relations however (the quantity of available money for instance, is fairly irregular with respect to temporal sub-abstraction). Here again, the characterization of the regularity of the relation allows to express a property between a couple of hierarchies in a simple and abstract manner, and simplifies the understanding of the knowledge base.

Following our example, there comes an interesting phenomenon: when Colombo decides to sell its car, the regularity is broken.

6.4. Rules that express irregularities

Regularity can also be used to characterize significant lacks of regularity. This is typically the case for the Selling or Buying events. In a selling event, the property of the object sold is transferred from the seller to the buyer. Note that this transfer of property is actually represented by a set of rules, and not a single rule : a rule to state that the seller owns the object before the selling event, a rule to state that the buyer will own it after the event, and

various rules to express properties concerning legal ownership, co-temporality (the buyer, the seller and the object must necessarily be at the same location at the same place), etc.

Here is (in a simplified syntax) a rule that would state the transfer of property of an object during a selling event: if anAgent performs aBuyingTransaction, the he/she will own the object after the transaction:

Rule: **OwnWhatYouBuy**
IF (allInstancesOf aTransaction Buying)
(occursIn aTransaction aSituation)
(nextSituation aSituation aSituation2)
(transactionObject aTransaction anObject)
(performedBy aTransaction anAgent)
THEN (holdsDuring anAgent owns anObject aSituation2)

Of course, this rule states some form of regularity of the world, in the etymological sense of the word: the rule describe regular selling events, as they occur most of the time. The important point to note here is that, according to our definition of regularity, the rule can also be seen as a regularity breaker. The rule justifies or explain the fact that the property owns is not always regular with respect to the temporal sub-abstraction relation. Stated in these terms the rule bears a more refined meaning that has more explanatory power than what is expressed in its standard representation.

7. Summary and conclusion

We are interested in the construction and maintenance of hierarchical semantic networks, and we have developed a model of hierarchies that generalizes taxonomic models and that supports a number of inferences that are more reliable than those related to inheritance (see section 8.2). We are also interested in the problem of generating argumentative documents in general, and for the case of non-specialized domains in particular, which necessitate a large volume of fine-grained common sense knowledge (see section 8.3.3). For these kinds of explorations, Cyc appealed to us as a realistic and convenient platform for testing our ideas and exploring new ones.

With regard to document generation, we ran into a number of problems early on. Some of these problems are not specific to Cyc and plague most sizeable knowledge bases, namely, the vocabulary problem: we have to know how to phrase the questions to get the right answers. The semantic brittleness that plagues traditional expert systems which Lenat calls “autistic”, has been replaced by a lexical brittleness. The latter is, in principle, manageable,

provided that we use a natural language interface to Cyc, that takes care of translating arbitrarily formulated requests into the appropriate terminology. It is interesting to note that one of the first applications of Cyc is natural language processing [Barnett et al., 1990]; let us hope that this mutual dependence does not turn out to be a deadly embrace! Because of its sheer size, Cyc also has problems of its own, which required some constraining design optimizations. In particular, selective reification (see section 8.5.1), combined with the notion of inference access levels (see section 8.5.2.2), led to a situation where users have to have a pretty good idea of the result, and of how to get it to ensure that they get the correct answers for their queries. That being said, we aren't ready yet to say that Cyc cannot support the generation of argumentative documents; we continue to experiment.

With regard to regularity, our explorations led to a number of interesting observations and venues, although different from the ones we anticipated. To the question "Does Cyc exhibit regularity", the answer is "yes". We also suspect that whenever its designers detected regularity patterns, they got rid of the actual values, and represented the regularity patterns with rules. For instance, a regularity rule expresses concisely and completely, what would have expressed, implicitly and incompletely, a diffuse set of Cyc assertions. We are currently exploring extensions to expansion (see section 8.2.3) that would enable us to reify concepts that would not have been reified otherwise. For example, it is possible to "create" the concept "EagleLike-Flying" or "SeagullLike-Flying" as the appropriate locomotion mode for the corresponding bird species, even if nothing explicit had been said about their locomotion modes[1].

Further, our analysis of regularity patterns and exceptions thereof led us to hypothesize that regularity could generally be used in tow ways: 1) to characterize the rules in the knowledge base that maintain/enforce it (e.g. rules managing the construction of temporal sub-abstractions, or the rules related to possession), but also 2) to characterize the rules of the knowledge base that violate/except it (e.g. rule stating the transfer of property after a sale). This last observation seems to suggest a venue for a problem related to argumentative document generation, namely, deciding which concepts, slots, or inference rules, are interesting in a proof trace, and are worth paraphrasing? We propose a simple rule based on regularity: the more regular a relation, the least interesting it is to paraphrase. Precisely, given two hierarchies and a link (property) or path connecting them, one of three cases may occur:

- The property is perfectly regular. This was the case for the cars and country of manufacture example. when such a regularity is considered essential (versus fortuitous), the property at hand is "uninteresting"; it expresses a tautology ("French cars are manufactured in France"). If the regularity is fortuitous, it may be interesting to the extent that it reveals an extrinsic aspect of the hierarchies at hand.

---

[1] which, according to the minimality principle (see section 8.5.1), would preclude the reification of such concepts.

- The property is utterly irregular: the hierarchies are not related. For example, there is no relationship between car models and the kinds of diseases that strike their buyers. At first glance, there is nothing to exploit in this case.
- The intermediate cases are more interesting: those for which a regularity is expressed and violated by some rules in the knowledge base. This was the case for the rules concerning property (possession) and temporal sub-abstractions. Exceptions to regularity reflect either missing concepts from the knowledge base (which expansion can fill in), or important properties of the world (e.g., property transfer following a sale transaction). The latter seem interesting to paraphrase.

Currently, we are exploring ways in which these and similarly simple rules may be used to develop "intelligent" document generation strategies, as discussed in section 8.3.3.

Acknowledgments: The work described in this chapter extended over a period of 8 years, and involved contributions from many people. Work on regularity and DC hierarchies was supervised by Prof. Roy Rada, currently with the University of Liverpool, while Mili was pursuing his doctoral studies at the George Washington University. Rachel Vincent inspired the journalistic example, and many more enriching thoughts and emotions. Profs. Gilles Gauthier, Robert Godin, Brigitte Kerhervé, Bernard Lefebvre, and Rokia Missaoui, all members of LARC, have: 1) graciously welcomed François Pachet during his post-doctoral stay at LARC, funded by the INRIA, 2) all, plus Prof. Jean-François Perrot of LAFORIA, have tirelessly steered the project back into line when it strayed, thanks to their questions, critiques, and suggestions, and 3) read and commented on different versions of this chapter; they are not responsible, however, for any errors, omissions, or plain crazy ideas that may have survived.

8. References

[Barnett et al., 1990] **Barnett J., Knight K., Mani I., et Rich E.**, «Knowledge and natural language processing,» *Communications of the ACM*, vol. 33, no: 8, pp. 50-71, August 1990.

[Brachman & Schmolze 1985] **Brachman R J. and Schmolze J. G.**, "An Overview of the KL-One Knowledge Representation System," *Cognitive Science*, vol. 9, pp. 171-216, 1985.

[Bush, 1945] **Bush, V.**, «As we may think,» *Atlantic Monthly*, (July 1945), pp. 101-108.

[Clancey, 1983] **Clancey, W.**, «The Epistemology of a rule-based expert system: a framework for explanation,» *Artificial Intelligence*, vol. 20, No 3 (1983), pp. 215-251.

[Collins & Loftus, 1975] **Collins A. M., Loftus E. F.**, «A spreading activation theory of semantic processing,», *Psychological review*, vol. 82, 1975, pp. 407-428.

[Conversations 1994] «The Knowledge Level in Expert Systems: Conversations and Commentary,» in *Perspectives in Artificial Intelligence*, Eds L. Steels and J. McDermott, Academic Press, 1994.

[Council 1988] **Council National Library and Information**, *Guidelines for Thesaurus Structure, Construction and Use*, Technical Report American National Standards Institute, 1988.

[CycBookReviews 1993] **«**Book reviews and response from Lenat & Guha,» *Artificial Intelligence*, vol. 61, pp. 37-181, 1993.

[First & al. 1985] **First M. B., Soffer L. J. and Miller R.A.**, "QUICK Index of Caduceus Knowledge: Using the Internist-1/Caduceus Knowledge Base as an Electronic Textbook of Medicine," *Computers and Biomedical Research*, vol. 18, pp. 137-165, 1985.

[Fisher 1987] **Fisher D. H.**, "Knowledge Acquisition via Incremental Concept Formation," *Machine Learning*, vol. 2, pp. 139-172, 1987.

[Guha 1991] **Guha R.V.**, *Contexts: A formalization and some applications.*, Technical Report ACT-CYC-423-91, MCC technical report, 1991.

[Guha & Lenat 1994] **Guha R. V. and Lenat D. B**, "Enabling Agents to Work Together," Communications of the ACM, vol. 37, pp. 127-142, 1994.

[Lenat & al. 1990] **Lenat D. B, Guha R. V., Pittman K. , Pratt D. and Shepherd M.**, "Cyc: Towards programs with common sense," *Communications of the ACM*, vol. 33, no: 8, pp. 30-49, August 1990.

[Lebowitz 1986] **Lebowitz M.**, "An Experiment in Intelligent Formation Systems: RESEARCHER," in *Intelligent Information Systems: Progress and Prospect*s, E. H. Limited, Ed., pp. 127-150, 1986.

[Lebowitz 1987] **Lebowitz M.**, "Experiments with Incremental Concept Formation: UNIMEM," *Machine Learning*, vol. 2, pp. 103-138, 1987.

[Lenat & Guha 1991] **Lenat D. and Guha R.V.**, *Ideas for Appling Cyc*, Technical Report ACT-CYC-407-91, MCC technical report, 1991.

[Lenat & Feigenbaum 1991] **Lenat D.B. and Feigenbaum E.A.**, "On the thresholds of knowledge," *Artificial Intelligence*, vol. 47, pp. 185-250, 1991.

[Lenat & Guha 1990] **Lenat D.B. and Guha R.V.**, *Building large knowledge-based systems. Representation and Inference in the Cyc project.* Addison-Wesley, 1990.

[Maida & Shapiro, 1982] **Maida A. S., Shapiro S. C.**, «Intensional concepts in propositional semantic networks,» *Cognitive Science*, vol. 6, 1982, pp. 291-330.

[Mayfield & Nicholas 1993] **Mayfield J. and Nicholas C.**, "SNITCH: augmenting hypertext documents with a semantic net," *International Journal of Intelligent and Cooperative Information Systems*, vol. 2, pp. 335-351, 1993.

[Mili 1988] **Mili H.**, *Building and Maintaining Hierarchical Semantic Nets*, Ph.D. thesis, George Washington University, Washington, D.C., 1988.

[Mili & Rada 1987] **Mili H. and Rada R.**, "Building a Knowledge Base for Information Retrieval," in *Proceedings Third Annual Expert Systems in Government Conference*, pp. 12-18, 1987.

[Mili & Rada 1988] **Mili H. and Rada R.**, «Merging Thesauri: Principles and Evaluation,» IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 10, pp. 204-220, 1988.

[Mili & Rada 1990a] **Mili H. and Rada R.**, "Generalizing Inheritance to Fuzzy Regularity," *IEEE Transactions on Systems, Man, and Cybernetics,* vol 20, No 5, Sept/Oct 1990, pp. 1184-1198.

[Mili & Rada, 1990b] **Mili H. et Rada, R.**, "Medical Expertext as Regularity in semantic nets," *Artifical Intelligence in Medicine*, vol. 2, pp. 217-229, 1990

[Mili & Rada 1992] **Mili H. and Rada R.**, "A Model of hierarchies Based on Graph Homomorphisms," *Computers and Mathematics with Applications*, vol. 23, pp. 343-361, 1992.

[Pinto et al. 1995] **Pinto N., Stephens L and Bonnel R.**, «Organizing domain theories for geographical reasoning using context,» in *Proceedings IJCAI'95 Workshop on "Modeling Context in Knowledge Representation and Reasoning*, Eds Patrick Brézillon, pp. 110-120, 1995.

[Quillian, 1968] **Quillian J. R.**, «Semantic memory,» in *Semantic Information Processing*, Eds Marvin Minsky, MIT Press, 1968.

[Rada 1989] **Rada R.**, "Writing and Reading Hypertext: An Overview," *Journal of the American Society of Information Science*, vol. 40, pp. 164-171, 1989.

[Rada 1990] **Rada R.**, "Hypertext writing and document reuse: the role of a semantic net," Electronic Publishing- Origination, Dissemination and Design, vol. 3, pp. 125-140, 1990.

[Rada & Barlow 1989] **Rada R. and Barlow J.**, "Expert Systems and Hypertext," *Knowledge* Engineering Review, vol. 3, pp. 285-301, 1989.

[Schmolze & Lipkis 1983] **Schmolze J. G. and Lipkis T. A.**, "Classification in the KL-One Knowledge Representation System," in *Proceedings Eigth Internationl Joint Conference on Artificial intelligence*, Karlsruhe, Germany, pp. 103-138, 1983.

[Smith 1991] **Smith B. C.**, "The owl and the electric encyclopedia," Artificial Intelligence, vol. 47, pp. 251-288, 1991.

[Streitz & al. 1989] **Streitz N.A., Hannemann J. and Thüring M.**, "From Ideas and Arguments to Hyperdocuments : Travelling through Activity Spaces," in *Proceedings ACM conference HYPERTEXT '89*, Pittsburgh, PA., pp. 343-364, 1989.

[Van Dam, 1988] **Van Dam, A.**, Hypertext'87 Keynote Address, *Communications of the ACM*, vol 31, No 7 (July 1988), pp. 887-895.

[Wang & al. 1991] **Wang W., Rada R. and Ghaoui C.**, "An Expertext Aid for Writing," in Proceedings World Congress on Expert Systems, Orlando, Florida, pp. 2767-2775, 1991.