Indexing Medical Reports in a Multimedia Environment : the RIME experimental approach

Catherine Berrut, Yves Chiaramella
Equipe Systèmes Intelligents de Recherche d'Informations
LGI-IMAG - BP53X - 38041 Grenoble Cedex - France
e.mail : berrut@imag.imag.fr, chiara@imag.imag.fr

abstract : *This paper focuses on the RIME system aimed to the indexing of medical reports in a multimedia environment. This particular application is viewed as representative of a large set of still unanswered needs of large communities of users : domain experts dealing with on-line specialized documentation such as software engineers, medical specialists and so on. In this application textual information appears as an interesting media to access related pictures in the data base. After the presentation of the application and a study of the particular corpus involved we define a semantic model for the documents which is based on a Conceptual Language. Then we detail the indexing process and its various linguistic components which perform the translation of every medical report according to this semantic model.*

I - Introduction

Every Information Retrieval System is based on the design of an underlying retrieval model which expresses a correspondance (or matching) function between documents content and user queries. An important component of this model is thus the document model which is aimed to provide a representation of the semantic content of the document in the system. The indexing process is devoted to the production, for every input document, of an internal representation of this semantic content according to a particular model which, in most industrial systems, is restricted to a set of keywords selected among a predefined list called the indexing language.

The level of the indexing language - ie its accuracy in the expression of concepts actually found in the documents - is clearly a key point considering the qualitative performances of the retrieval systems, in terms of recall and precision. Many attempts have been made towards an improvement of such document models, for instance in considering noun phrases instead of keywords [CHIA86] and furthermore in deriving semantic interpretations of their meaning [SPAR79], [CROF86], [SMEA87]. Applied to what we call open corpuses - ie texts with unrestricted vocabulary and consequently unrestricted semantic domain - such approaches are clearly difficult to apply due to the overwhelming problems related to linguistics and semantics : they are at the moment restricted to more specific applications. On the other hand, both linguistic and semantic problems may be solved at a reasonable cost in applications where texts present particular linguistic attributes, and are related to a well identified semantic domain.

We have to point out here that the theoretical and practical interest of such applications, while not very familiar in the area of Information Retrieval, is very important. Considering their practical interest and their economic impact there are indeed considerable needs in information retrieval in strategic domains like software engineering, space engineering, medicine for example. In these areas, the users need "access by content" facilities to on-line technical or scientific documents such as reference manuals, specifications, technical documentation of products, reports and so on. These functionalities are complementary to classical browsing techniques which are currently provided by multimedia data bases or office information systems. About the theoretical interest, the design of such elaborated retrieval models and their application to real size problems is in our opinion among the best ways to assert what really might be transferable to more classical applications through the design of enhanced

retrieval models. It should also be emphazised that the users of such systems are most often experts in the particular domain : a particular orientation is thus given towards high precision for the expected answers.

In this paper we focus on a particular application of this kind : the RIME (for **R**echerche d'**I**nformations **ME**dicales) retrieval system which is aimed to the retrieval of iconographic documents in a medical environment [BERR88a]. The iconographic corpus is a collection of X-ray pictures, each of them being related in a one-to-one relationship with a medical report which describes in natural language its content and the medical interpretation of what the radiologist has observed on that picture. Both iconographic and textual information are stored within a multimedia system. These aspects have been investigated in [MUNO87] who implemented a prototype based on the TIGRE multimedia DBMD [LOPE83] and later on the ORACLE DBMS. The prototype allowed to access these data using SQL-like queries. In this particular context it is clear that information retrieval techniques applied to the set of medical reports is an indirect way to retrieve the associated X-ray pictures in the multimedia system : every report is in fact an indexing of the corresponding picture. A proper indexing of the medical reports may allow to answer queries like *select all the X-ray pictures where an observed lung opacity is related to lung cancer.*

We present here the semantic model (the indexing language) for the medical reports and the corresponding indexing process which have been designed for this application, and which involves a deep understanding of the technical language of the radiologists. This has been made possible due to the restricted domain addressed by the documents and the linguistic properties of this highly technical language. Our claim is that this approach may be successfully applied to many other fields which present the same characteristics. We thus try to design tools in order to facilitate such transfers from one domain to an other. Though still in development in the context of the IOTA project, the indexing component of the RIME project has by now reached a good level of achievement which allows us to present here its main attributes and preliminary results.

II - The semantic model

The aim of this chapter is to design the indexing language of the RIME information retrieval system.

2.1. presentation

Medical reports most often are short hand-written documents, usually less than one page long, which are produced by specialists while they investigate medical data. They usually combine external attributes such as patient's names and dates, examination data about the special techniques used for the examination, observations and, most often, a proposed diagnosis. These three last kinds of information constitute the very content of the medical reports, in contrast with external attributes which describe the context of the examination. While external attributes are usually described in a conventional format (i.e. using specific predefined fields in the document) and are well adapted to a relational model, the content is expressed in natural language and needs a more sophisticated and flexible model.

Due to the intrinsic nature of these documents and the circumstances of their elaboration, they are both highly technical and written in a "speciality language", that means :

- the basic vocabulary is mainly made of technical terms related to the technology used for the examinations, and of medical terms used to describe observations and to issue diagnosis ;
- concise and direct style most often lead to rather simplified linguistic analysis.

These two fundamental characteristics of the textual information we have to process may allow automatic deep understanding of medical reports which, in turn, may allow the design of high-level information retrieval access techniques. In a classical approach, deep understanding means the possibility to translate natural language information in an elaborate semantic model, and high-level information retrieval techniques means the possibility to issue "content oriented" queries.

2.2. main principles

According to the properties of these documents, our basic hypothesis is that we are able to design an automatic indexing method which allows 'text content' research, that means deep understanding of these medical documents through the use of a complex indexing language. Complex indexing languages enable pre-coordinate indexing, that means that they enable the extraction of coordinated index terms (in these methods, the index terms are concepts coordinated with each other during the indexing time), in regards to post-coordinated methods (in those methods, the

index terms are single concepts which are coordinated at the time of searching) [RIJS79], [SALT80].

It seems clear that talking about "text content", even about such specific documents as medical reports is a complex issue : what happens in our mind when we read a simple noun phrase like "lung opacity"? We may stay to this factual information, or think about "lung" as a component of the human body, or think about the nature of this "opacity" and so on. The main problem while designing a semantic model for documents is thus to determine a limit in the process of understanding natural language sentences. In this particular application this has been done according to the level of precision expected from the retrieval process : the semantic model corresponds to factual knowledge extracted from natural language sentences and its granularity (its level of precision) has been carefully defined with the assistance of radiologists who are potential users of the retrieval system. As an example the term "lung" is kept as a basic concept in the model, with no information such as "element of the breath system". This kind of knowledge which is clearly a semantic interpretation of "lung" is stored within a thesaurus and may be accessed for proper inference during query processing.

Given this convention the question now arises about the knowledge representation choosen in this context. The model proposed here is derived from the notion of *conceptual dependency* introduced by SCHANK [SCHA80], [SCHA81] which has been enriched here through the definition of a *Conceptual Language* which gives a very convenient way to control and further to use - at retrieval time - the concepts. The representation of every medical report according to this model is called a *Conceptual Report.*
The design of the model is based on the following principles :

a) the notion of dependancy is expressed through binary tree structures :

- the non-terminal nodes correspond to semantic operators which explicitate particular relationships between lower-level concepts (i.e. subtrees).
Example : the binary operator **due-to** establishes a causality relationship between its two operands.

- the terminal nodes correspond to medical or technical terms which are simple words or compound words in common use in the domain (they in fact define the semantic domain of the application).
Example : "liver", "lung", "opacity"..

b) every sentence of a medical report is translated in such a tree which constitutes the proposed representation of the meaning of this sentence. Hence the conceptual model of a medical report is a set of such trees.
Example : the following expression is the prefixed notation of a simple binary tree which corresponds to the interpretation of the sentence "the lung opacity" : **[bears-on**, opacity, lung].

c) every tree is built according to a formal model which is defined by a grammar. The language which corresponds to this grammar is called the Conceptual Language (or the target language when considering the interpretation process from the natural language to the conceptual model). We call Conceptual Report (CR) the representation of a medical report according to the semantic model.

Thus all the indexing process of the medical report may be seen as a *translation* from natural language sentences to Conceptual Language elements.

The organization of knowledge imbedded in medical reports clearly presents several levels of organization. As mentioned before, a medical report usually contains information about the examination, the observations made and possibly a diagnosis. Each of these elements may in turn be refined in terms of subnotions such as signs, localization and so on, the lowest level being elementary medical facts or knowledge. This obviously suggests a hierarchical conception of the model which fits well with the tree structures defined above. Moreover this hierarchy may be defined through a context free grammar where the metasymbols stand for intermediate-level concepts (such as signs, observations, diagnosis), and terminal symbols stand for atomic concepts (or self-defined concepts). The rules of the grammar provide the definition of the intermediate concepts (see examples below) and specify the possible structures of the trees. The correspondance between metasymbols and intermediate-level concepts is very important because it allows to retrieve these concepts in conceptual reports, using syntactical properties (hence the implicit knowledge to which they correspond may be made explicit).

2.3. the conceptual language

The knowledge imbedded in medical reports presents several levels, or meta-notions, which

correspond to a hierarchical view of their content. Starting from top-level notions, we mentioned before that a medical report usually contains information about the examination, the observations made by the radiologist, and possibly a diagnosis. These notions are interesting by themselves for example if we want to further retrieve all the reports having a *diagnosis* related to cancer, or all the reports involving *examination* of lung. Going down to more detailed notions we can observe that a diagnosis may be the identification of a *lesion* (such as a tumor), or that an observation implies the identification of *clinical signs* (such as an opacity). Again these subnotions are interesting to consider for retrieving reports related to *clinical signs* observed on lungs or *lesions* related to lung. At the lowest level, we find the atomic concepts already described in the section above. In the model they constitute particular instances of upper level meta-notions : for example, "opacity" is an instance of the meta-notion "clinical sign" (which will be further abbreviated into "sign").

Due to the obvious interest of this hierarchical definition of the domain knowledge of medical reports for retrieval purposes, we found very convenient to express all these relationships using a formal grammar. Moreover this formal model also gives us a convenient way to implement the various specifications of the semantic model presented in the previous section as may be seen below :

i. the conceptual language is defined by a *context free grammar*. Like every grammar of this kind, it is defined by a *terminal vocabulary* (set of terminal symbols) , a set of *non-terminal symbols* (or meta-symbols), a *set of rules*, and a particular meta-symbol called the *initial symbol*.

ii. the terminal vocabulary contains the set of the atomic concepts of the model and the set of the semantic operators of the model (see i. in section 2.1). Every terminal symbol will be typed in lower-case character, the semantic operators being between brackets to avoid confusions. Example : "lung", "**due-to**"...

iii. the non-treminal vocabulary contains the set of all the meta-notions of the semantic model (see the discussion above). Every non-terminal symbol will be typed in upper-case character. Example : "SIGN", "DIAGNOSIS"...

iv. the initial symbol is "CR" which stands for "Conceptual Report", and which corresponds to the highest level meta-notion in the model.

v. the rules are context free rules : the left part of each rule contains a unique meta-symbol. In the examples below they will be presented using the BNF format :

<left part> ::= <right part>

vi. every rule may derive only a binary structure. This means that the right part of every rule contains at most two meta-symbols and an operator, or only one meta-symbol. This last case corresponds to the definition of equivalences among meta-notions.

Examples :

OBSERVATION ::= [**shown-by**, SIGN, EXAMINATION] | SIGN

which states that an observation may be expressed either by the connection of a SIGN and an EXAMINATION by the semantic operator **shown-by**, or by a SIGN alone (the symbol "|" in the right part defining an alternative definition of the rule).

DIAGNOSIS ::= LESION

which states that a diagnosis may be equivalent to the identification of a lesion.

vii. the grammar is not ambiguous. This means that there is not set of rules having the same right part and different left parts.

viii. the terminal rules of the grammar (with the right part reduced to a terminal symbol) define the relationship between atomic concepts and meta-notions.

Example : SIGN ::= opacity | deviation | ...

Establishing a connection with a well known terminology in artificial intelligence, we may say that the set of terminal rules define *semantic classes* among atomic concepts. Considering the current example above, we may say that "opacity", "deviation" belong to the same semantic class of concepts called SIGN.

There is no terminal rule involving a semantic operator in its right part because we have not considered any notion of semantic class of operators until now in the model. This should correspond to classes of operators having a common behaviour in upper level rules. This classification appears to be possible in the model (see examples below with the rules related to LESION) but does not seem very interesting at least in the first phasis of the definition of the model because it would somewhat increase its complexity.

Given this definition of the grammar we can say that a Conceptual Report is a set of words from the corresponding language. Every sentence of a

medical report will be coded as a word of this language which in turn may be seen as a binary tree the definition of which fitting with the specifications given in section 2.1. Considering the content of a Conceptual Report, we can say that every subtree appearing in its definition corresponds to a meta-notion - or high-level concept - which may be uniquely identified and hence retrieved. Though not being the main topic of this paper we shall give some hints about the planned retrieval process in section 3 below.

2.4. the application of the Conceptual Language to the medical reports

As said before this work has been done in close relationship with experts who could provide on one side an accurate and consistent definition of the domain and on the other hand a clear idea of the typology of the users' potential queries.
Considering the medical reports we quickly identified the various typical components of these documents and thus defined the hierarchy of meta-notions in a top-down way. The examples of rules given below illustrate the three main levels so identified in this context.

a) the first level expresses that a report is made of one or several sentences, and gives the formal definition of higher-level concepts related to the main components of a medical report (conceptual report, observations, diagnosis) :

CR	::=	OBSERVATION
CR	::=	DIAGNOSIS
CR	::=	[allows_to_deduce, OBSERVATION, DIAGNOSIS]

these rules define a CR (conceptual report) as possibly made of an observation, a diagnosis or an observation and a related diagnosis (the relationship being made explicit by the semantic operator **allows_to_deduce**).

b) the second level defines basic notions such as OBSERVATION, DIAGNOSIS and the associated sub-notions.

OBSERVATION	::=	[due_to, OBSERVATION, OBSERVATION]
OBSERVATION	::=	[shown_by, SIGN, EXAM]
OBSERVATION	::=	SIGN
DIAGNOSIS	::=	[and, DIAGNOSIS, DIAGNOSIS]
DIAGNOSIS	::=	LESION
SIGNE	::=	[has_for_value, SIGN, QUAL]
LESION	::=	[bears_on, LESION, LOC]

OBSERVATION ::= **[due_to,** OBSERVATION, OBSERVATION]
this rule expresses that observations may be interdependant

OBSERVATION ::= **[shown_by,** SIGN, EXAMEN] | SIGN
this rule defines an observation as a sign revealed by an examination, or a sign alone. A sign is an observable entity, such as a volume, an area.

DIAGNOSIS ::= **[and,** DIAGNOSIS, DIAGNOSIS]
this rule defines a diagnosis as a possible combination of several sub-diagnosis.

DIAGNOSIS ::= LESION
this rule defines a diagnosis as the identification of a particular lesion.

SIGN ::= **[has_for_value,** SIGN, QUAL]
this rule defines a sign as being possibly related to a qualifier (such as augmentation, diminution).

LESION ::= **[bears_on,** LESION, LOC]
this rule defines a lesion as being related to a particular localization.

c) The third level contains pre-terminal and terminal rules, which correspond to the lowest level concepts in the model. The set of terminal symbols is defined in a dictionary (see below); the sub-classes of terminal symbols (like SIGN...) are determined by assigning each term in the dictionary a semantic class which is the corresponding metasymbol (as an example, the term "augmentation" will be assigned the semantic class "sign" in the dictionary). Being associated to metasymbols in the grammar, the semantic classes assigned to terms are used while generating the semantic interpretation of sentences. The rules in fact control which classes may be combined to derive a correct interpretation. This approach is very similar to Semantic Grammars introduced by Fillmore [FILL68].

SIGNE	::=	{t ∈ Vt / cat_sém(t) = sign }
LESION	::=	{t ∈ Vt / cat_sém(t) = lesion }

SIGN stands for "a term whose semantic class is sign" (like "volume"..).
LESION stands for "a term whose semantic class is lesion" (like "tumor", "cancer"..).

Of course this gives only a partial idea of the actual semantic model which is made of about 60

rules; as an indication, there are about 10 different semantic operators and the dictionary is organized in about 20 semantic classes.

III. The retrieval process

The retrieval component is still under development at this moment. It accepts natural language queries which are interpreted in the same way as medical report sentences. Because documents and thus related pictures in the system are accessible through external attributes such as patient names, dates and so on and content attributes expressed according to the Conceptual Language, the query interpretor has to be able to find out within the natural language query these two kinds of search criterias and to combine them properly in an internal representation. This kind of analysis has already been succesfully implemented in IOTA [DEFU86]. The only difference in this application lies in the interpretation of the content attributes which has to be done according to the specific model of RIME. Because retrieving using external attributes may be performed using classical access techniques in database systems we shall focus here on access involving content attributes.
Basically the content search criteria of the query is a boolean expression of concepts expressed according to the Conceptual Language. Every document satisfying the query
Retrieving documents from a query may be seen as a matching operation between two sets of trees : the first corresponds to the representation of documents content, the second corresponds to the interpretation of the query.

IV - The indexing process
4.1. principles

The RIME indexing process consists in defining a function which is able to translate a medical report into its corresponding conceptual report, that is its representation according to RIME semantic model.

From a general point of view, this process might be seen like a classical automatic translation, which aim is to translate one language to another considering all the appearing linguistic phenomenas and ambiguities. Actually, we do not have here to deal with all these linguistic problems that appear in general automatic translation :

- on one part, the language which is used in the medical documents is a speciality language. That means that it of course uses general properties of a natural language, which is French in our application. But it works on a subset of words through a specialised vocabulary, and it also uses limited syntax (complex nominal syntagms generally) and semantics (close universe of discourse, limited to the radiological medical domain). Consequently, such a language does not lead to all the linguistic problems that a general natural language might have. This is, comparing with general automatic translation problems, our first important simplification ;
- on the other part, the indexing process has to generate reports written in the conceptual reports. Because of the well formalization of the conceptual language, its expression and generation do not lead to problems as ambiguous and complicated as those found in natural languages generation. This, in regards to general translation, indicates our second important simplification of the problem.

Considering these two aspects of the problem, we first studied both the reports language and the conceptual language, in order to be able to give a list of the linguistic phenomenas which have to be managed to process RIME indexing (see section 4.2). Secondly we designed the architecture of the indexing process, and the indexing linguistic process which are able to handle the needed linguistic phenomenas (see section 4.3).

We shall not here in this paper give a detailed description of each linguistic process we built for RIME, but we only would like to give an overview of them. Actually we would like here to show the necessary minimum to prove that our approach of deep inderstanding during indexing is a manageable approach. For a detailed and formal description of this process, see [BERR88b].

4.2. linguistic phenomenas
4.2.1. introduction

We do not want in this chapter to specify any linguistic process of RIME indexing (morphology, syntax, semantic). We just want to describe the different linguistic tasks that have to be processed to transfom the text from a set of single words to its conceptual representation.

Actually any natural language processing has:
- to recognize each word, to establish its virtual attributes and to contextually deduce its actual attributes, that is what we call **infra-structural tasks**;
- to agregate words together to build structures, and to nominate structures and sub-structures, that is what we call **intra-structural tasks**;
- to deduce implicit links between structures, that is what we call **inter-structural tasks**.

While solving these tasks, we also have to deal

with their associate problems (ambiguities for example). We show here which of these different tasks are needed in RIME, and also their associate problems.

To define these different linguistic tasks, we need notions through which our structures are represented. First of all, we define
- a simple word as a contiguous sequence of characters (letters and numbers in Latin languages);
- non-imperative separators as the set {blank, quote, hyphen} i.e. {" ","'","-"};
- a word as a meaningful and contiguous sequence of simple words separated from each other by non-imperative separators.

4.2.2. infra-structural tasks

At this step, words are considered as individual entities without link with each other.

4.2.2.1. infra-structural tasks

a) **identification of the words**
That means:
- to isolate each simple word;
- to deduce from the set of simple words the set of words according to the universe of discourse.

Example : the sentence *extension ganglionnaire médiastinale au niveau du groupe de la bifurcation de la chaîne para-trachéale droite* is compound of 15 simple words, which correspond to 13 words, because the words *au niveau du* constitute a unique word in our medical corpus (that of course would probably not be the case in another universe of discourse);

b) **identification of the virtual attributes** of each word, i.e. the set of its potential morphologic, syntactic and semantic attributes.
Example : the word *ganglionnaire* has for morpho-syntactic virtual attributes *masc,sing* or *fem,sing*, for virtual syntactic attributes its syntactic category *adjq*, and for virtual semantic attributes: *ganglion, const-org*. These three subsets mean that the word may be masculine or feminine, depending of its context, that it is an adjective, and that its virtual reference is in our universe *ganglion* which is an organism-constituant.

c) **deduction of the actual attributes** of a word from its set of virtual attributes and from the correspondant sets of attributes of its context.
In the previous example, *extension* in the right context of the word *ganglionnaire* which virtual morpho-syntactic attributes are *(fem,sing)* allows us to deduce that the actual morpho-syntactic attributes of the word *pulmonaire* are *(fem,sing)* because of the French agreement laws between nouns and adjectives.

4.2.2.2. infra-structural problems

The set of actual attributes of a word may be not completely determined at this infra-structural level. There actually exist three possibilities :
- morphologic ambiguities like *temps de coagulation dangereux,* where we are not able to decide whether we speak of one or more *temps*, because the singular and the plural of the words *temps* and *dangereux* are identical;
- syntactic ambiguities (homographies) like *porte*, where we are not able to decide whether we speak of the substantive meaning *door* or of the verb meaning *to bear*.

We do not in RIME consider pure polysemies, which in our specific domain do not exist.

4.2.3. intra-structural tasks

The aim of the intra-structural tasks is to agregate words together, to build structures which can be either syntactic or semantic.

4.2.3.1. these intra-structural tasks are:

- to **build** syntactic or semantic structures;
- to **nominate** syntactic or semantic structures.

4.2.3.2. intra-structural problems

The building of structures leads to two kinds of ambiguities:

- **attachment** : the problem here is to identify the links of immediate dependencies in structures compound of at least three constituants. Let us consider the structure made of (X Y Z) where Z is not independant and has to be linked to either X or Y, hence there exist at most three potential configurations of links between X,Y,and Z: (X(Y Z)) or (X Y(Z)) or (X(Y)(Z)). For example, *confirmation d'une hypertrophie de densité tissulaire* may be analyzed as *(confirmation (d'une hypertrophie de densité tissulaire)), (confirmation d'une hypertrophie (de densité tissulaire))*, and *(confirmation (d'une hypertrophie) (de densité tissulaire))*.

- **closure** : let us consider a structure made of at least three constituants (X,Y,Z) where X may be independant, that means that X may be not linked,

or linked to Z though in any case Y is linked to Z. For example, *le cancer et la tumeur du poumon* which may be analyzed as *(le cancer) et (la tumeur du poumon)* or as *(le cancer et la tumeur) du poumon.*

4.2.4. inter-structural tasks

The aim of these inter-structural tasks is to deduce implicit links between structures. This may happen through transformations between structures because of their proximity of another structure.

4.2.4.1. list of transformations

a) **deletion** of a part of a structure

We only treat the deletion of repeated constituants in the following structures:

- the **coordinate** structures

For example, in the sentence *opacité pulmonaire en projection du lobe supérieur droit et d'aspect alvéolaire avec signe de nécrose*, a part of the second member of the coordinate structure *opacité pulmonaire* has been deleted through the use of the coordination conjunction *et* ;

- deletion due to **possessive** adjectives

For example, in the sentence *La lobaire supérieure droite présente une amputation au niveau de sa lumière*, there is a deletion of *de la lumière de la lobaire supérieure droite* through the use of *sa* ;

- the **comparative** structures

Example : *la tumeur est plus grosse que sur la radio du 30.10.87* which could be re-established as a double-structure sequence *la tumeur est de **taille** plus grosse que la **taille de la tumeur** de la radio du 30.10.87* ;

b) **anaphoras**

- **Nominal anaphoras** through repetition of constituants which actual attributes are identical.

Example : ***opacités** alvéolaires en projection de la lingula, qui est le siège d'une rétraction modérée.; ces **opacités**,...*

In RIME, these anaphoras are recognized if and only if they are introduced by demonstrative adjectives;

- **Nominal anaphoras** through inclusion of the semantic virtual attributes of a first constituant in a second one.

Example : *Aspect TDM en faveur d'**un cancer** de siège périphérique. Extension de **la lésion** au niveau médiastinal et pédiculaire.*

- **Pronominal anaphoras**

Example : *mise en évidence au niveau de la loge latéro trachéale droite d'**opacités** évocatrices d'**hypertrophies ganglionnaires**. **Elles** mesurent environ 10 mm de diamètre*

c) **reach of operators**

This happens for example on negative sentences. Example : *Il **n'y a pas** de cancer du poumon* where the negation can reach either *cancer* or *cancer du poumon.*

4.2.4.2. inter-structural problems :

- to identify the traces of these transformations, through syntactic phenomenas of the text surface; that means to identify the terms through which an anaphora might be detected;
- to find in the context the set of potential referents;
- to decide of one solution from this set;
- to re-establish as far as possible the "previous to modification" state of the structure, that means to create a link between the structures.

4.3. the indexing architecture

Considering all these linguistic phenomenas, we designed different linguistic process which are able to handle them : a morphological process, a syntactic process and a semantic process (see figure 1). Each of these process is specified according to the linguistic phenomenas it has to deal with. For example, the morphological process has to extract each word of the text, and to deduce the virtual attributes of each word.
Another fondamental study in RIME was also to design a system architecture managing these different linguistic process. We decided to build a system in which each linguistic process was independant to each other. Such a decision implies that we had to build a process called the cooperation process which is aimed to manage the translation, and to store the entire memory of the whole translation. See figure 2 the architecture of RIME indexing system.

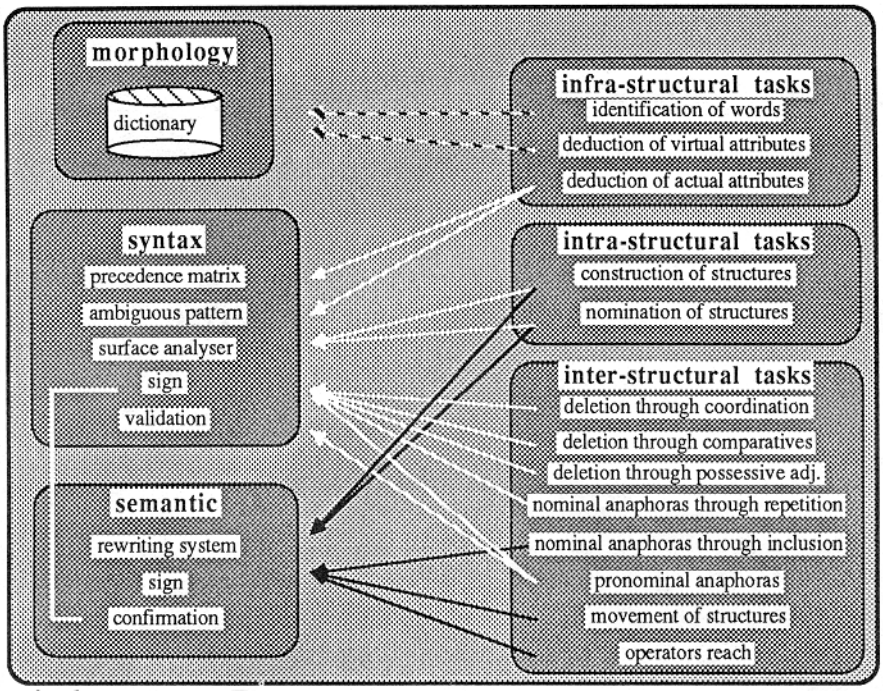
4.4. the lexicon

The first linguistic tool we need is a lexicon, in which we store the morphological, syntactic and semantic attributes of each stored word.

4.4.1. morpho-syntactic attributes

We need for each word morpho-syntactic informations. These informations are divided in 2 parts :

figure 1 : TASKS DISTRIBUTION IN RIME



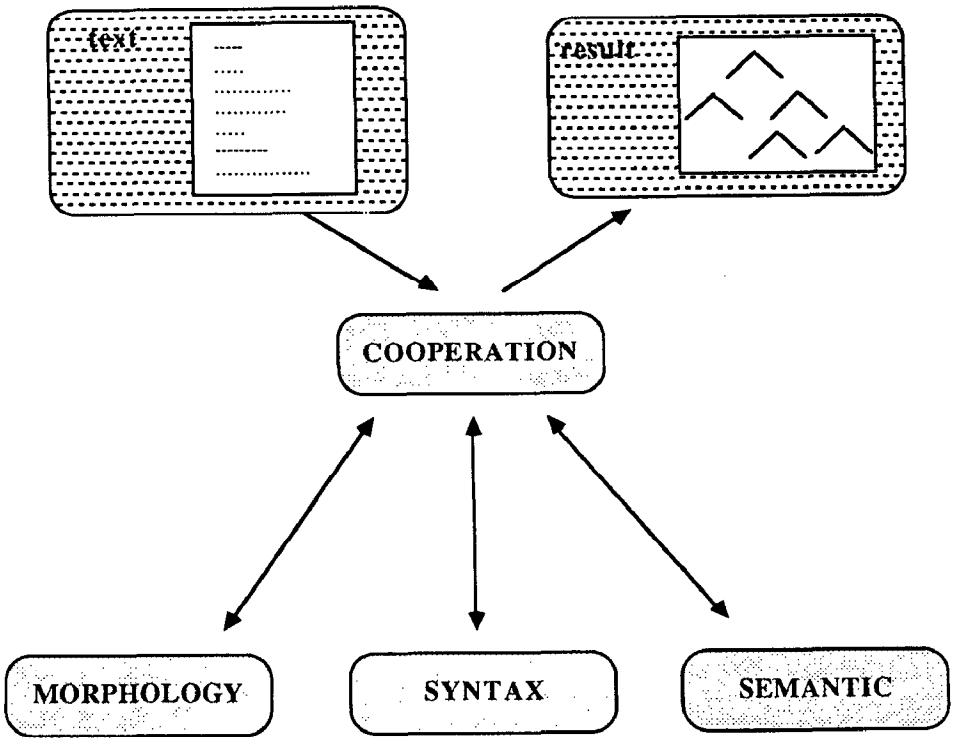
- the grammatical category. For example *lung* is a common substantive ;
- the grammatical values. For example, *lung (poumon* in French) is masculin and singular.

4.4.2. semantic attributes

We also need semantic informations for each stored word. This information is also divided in 2 parts :

- the semantic category, which corresponds to a grammar terminal of the Conecptual Language. For example *lung* is a *const_org* (organism constituant) ;
- the semantic feature, which corresponds to the word interpration according to the grammar. For example *lung* corresponds to itself, but

figure 2 : RIME architecture



leucemia corresponds to *[bears_on, cancer, blood].*

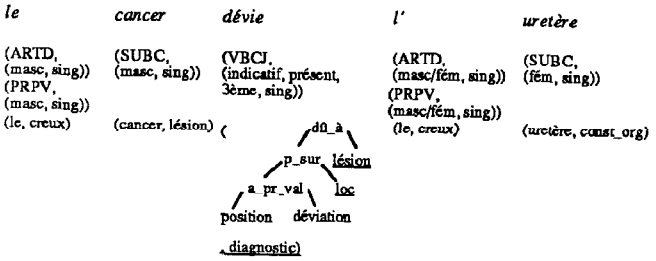
4.5. the different linguistic process

We will show the different linguistic process through an example.
Let the sentence *le cancer dévie l'uretère (the cancer deviates the ureter)* be translated.

4.5.1. cooperation and morphology

The first task of the cooperation process is to give this sentence to the morphology process. The morphology processes this sentence first to extract its words and secondly the virtual attributes of each word through a lexicon access.

Morphology result :



4.5.2. cooperation and syntax

The cooperation process transforms this result to give it to the syntax.

Syntax entry :

le	*cancer*	*dévie*	*l'*	*uretère*
(ARTD, (masc, sing)) (PRPV, (masc, sing))	(SUBC, (masc, sing))	(VBCJ, (indicatif, présent, 3ème, sing))	(ARTD, (masc/fém, sing)) (PRPV, (masc/fém, sing))	(SUBC, (fém, sing))

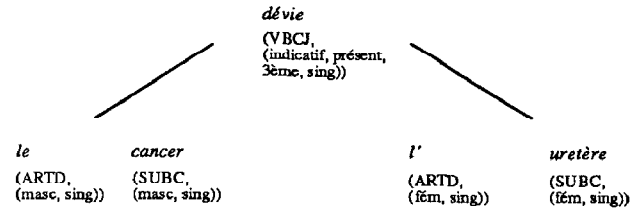
The syntactic process treats the entry in 2 steps :

- first, it processes intra-structural tasks : it simplifies the syntactic possibility through a double syntactic filter (a precedence matrix, and a list of ambiguous schemas) [BERR86] ;

le	*cancer*	*dévie*	*l'*	*uretère*
(ARTD, (masc, sing))	(SUBC, (masc, sing))	(VBCJ, (indicatif, présent, 3ème, sing))	(ARTD, (fém, sing))	(SUBC, (fém, sing))

- secondly, it processes inter-structural tasks : it builts a syntactic tree in which nominal, prepositional and verbal syntagms appear.

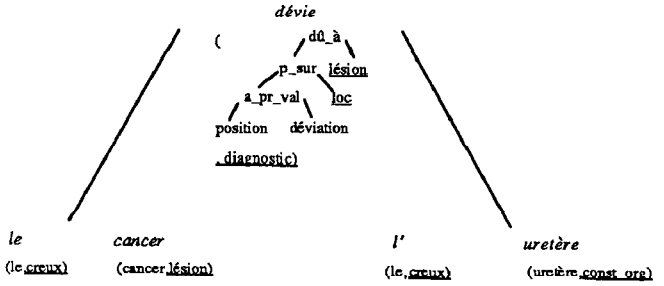
Result :



4.5.3. cooperation and semantic

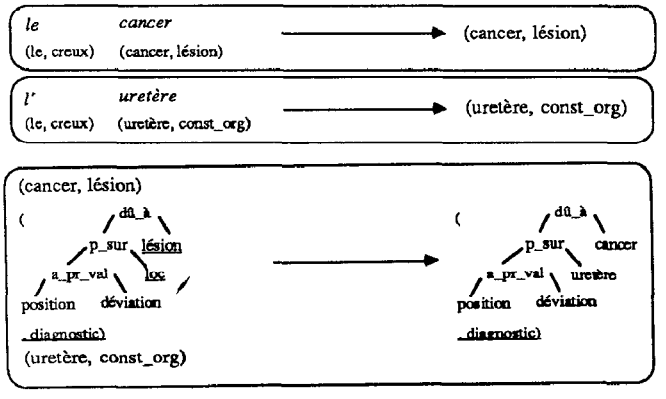
The cooperation process treats this syntactic result and transforms it to give it to the semantic process.

Semantic entry :



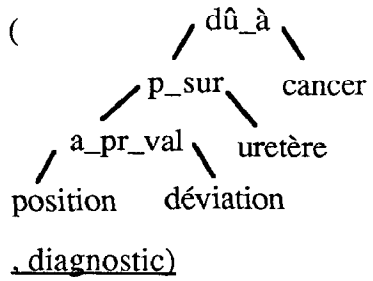
The semantic treats the entry through its 2 components : the semantic envelop and the semantic kernel. The semantic envelop constitutes the syntax/semantic interface (the semantic envelop manages the intra-structural tasks : resolution of anaphoras, resolution of deletion through possessive adjectives, ...), the semantic kernel represents the pure semantic part of the semantic process (the semantic kernel manages the inter-structural tasks : it builts semantic structures according to the Conceptual Language).

semantic envelopp call and semantic kernel answer



The semantic result is considered as the final result of the translation.

Final result :



V - Conclusion

The first experimentations in Prolog on a Sun workstation have led to very encouraging results, which demonstrate the validity of the approach. But they also clearly show that the effectiveness of the process has to be improved; this mainly concerns the design of updating interfaces for the dictionary , the syntactical and semantic analysers.

References

[BERR86] C. BERRUT, P. PALMER
Solving grammatical ambiguities within a surface syntactical parser for automatic indexing
ACM-SIGIR86, Pise, 1986.

[BERR88a] C. BERRUT, P. CINQUIN
Natural language understanding of medical reports
IFIP-IMIA International Working Conference on Computerized natural language processing for knowledge representation, Geneva (Switzerland), 1988.

[BERR88b] C.BERRUT
Une méthode d'indexation fondée sur l'analyse sémantique de documents spécialisés. Le prototype RIME et son application à un corpus médical
PhD thesis, december 1988, Grenoble University (France).

[CHIA86] Y. CHIARAMELLA, B. DEFUDE, D. KERKOUBA, M.F. BRUANDET
IOTA : a prototype of an information retrieval system
ACM-SIGIR, Pisa (Italy), 1986.

[CROF86] W.B. CROFT, D.D. LEWIS
An approach to natural language processing for document retrieval
ACM-SIGIR, New Orleans (USA), 1987.

[DEFU86] B.DEFUDE
Etude et réalisation d'un système intelligent de recherche d'informations : le prototype IOTA
PhD thesis, july 1986, Grenoble University (France).

[FILL68] C.J. FILLMORE
The Case for Case, Universals
Linguistic Theory, 1-88, E. BACH and R.T. HARMS, Molt Rinehart and Wiston, New York.

[LOPE83] M. Lopez, J. Palazzo, F. Velez
The TIGRE data model
IMAG et Centre de Recherche BULL, TIGRE Research report, Grenoble (France), November 1983.

[MUNO87] G. MUNOZ BACA
Stockage et exploitation de dossiers médicaux multimédia au moyen d'une base de données généralisée.
PhD thesis, july 1987, Grenoble University (France).

[RIJS79] C.J. Van RIJSBERGEN
Information retrieval
Second edition, Buttersworth London, 1979.

[SALT80] G. SALTON
Automatic information retrieval
Computer, vol 13, n°9, 1980.

[SAGE78] N. SAGER
Natural language information formatting : the automatic conversion of texts to a structured data base
Advanced in computers, vol 17, 1978.

[SCHA80] R. SCHANK
Language and memory
Cognitive Science, vol. 4, pp. 243-284.

[SCHA81] R. SCHANK
Representing Meaning : An Artificial Intelligence Perspective
Cognitive Science Technical Report 1, Yale University, Avril 1981.

[SMEA87] A. SMEATON
Using parsing of natural language as part of document retrieval
PhD Thesis, Glasgow, 1987.

[SPAR79] K. SPARCK JONES
Problems in the representation of meaning in Information Retrieval
ASLIB Informatics Group & British Computer Society, Information Retrieval Group, 1979.