# Structuration and Acquisition of Medical Knowledge

## Using UMLS in the Conceptual Graph Formalism*

F. Volot, M.D.[1], P. Zweigenbaum, PhD.[2], B. Bachimont, PhD.[2], M. Ben Said, M.D., M.S.[2], J. Bouaud, PhD.[2], M. Fieschi, M.D., PhD.[1], J.F. Boisvieux, M.D., PhD.[2]

[1] Service de l'Information Médicale, Hôpital de la Timone, boulevard Jean Moulin, F-13385 Marseille Cedex 05 - France

[2] DIAM, INSERM U194 & Service d'Informatique Médicale AP-HP, 91, boulevard de l'Hôpital, F-75634 Paris Cedex 13 - France

*The use of a taxonomy, such as the concept type lattice (CTL) of Conceptual Graphs, is a central structuring piece in a knowledge-based system. The knowledge it contains is constantly used by the system, and its structure provides a guide for the acquisition of other pieces of knowledge. We show how UMLS can be used as a knowledge resource to build a CTL and how the CTL can help the process of acquisition for other kinds of knowledge. We illustrate this method in the context of the* MENELAS *natural language understanding project.*

## INTRODUCTION

Taxonomies and hierarchies are widely used in medical knowledge based systems [1, 2]. Some knowledge representation formalisms indeed give a central place to *is-a* hierarchies. In particular, the Conceptual Graph (CG) formalism [3] is organized around a central Concept Type Lattice (CTL). The CTL provides the basic concept vocabulary that all domain descriptions will be built of. It is shared by the different knowledge sources of the system. Its hierarchical structure determines the valid operations on knowledge representations. The issue is then to build this CTL.

UMLS [4] is a repository of a large body of medical knowledge, and its semantic network [1] can be used as a resource for this purpose. However, UMLS is still limited in its coverage [5, 6], and must be adapted to specific needs. The CTL built using UMLS can then further help the process of acquisition for other kinds of knowledge. We illustrate this method in the context of the MENELAS natural language understanding project.

This paper is organized as follows. First, we start with a brief recall of basic notions and related literature. Second, we describe the structuration of knowledge in MENELAS, and the way it rests on a CTL. Third, we show how UMLS has been used to initiate CTL building, then illustrate how the CTL structures the knowledge acquisition process. Finally, we discuss the advantages and further issues associated with this approach.

*This work has been partly supported by the European Community project MENELAS (AIM 2023).

## BACKGROUND

### Taxonomies and Hierarchies in Medical Knowledge Based-Systems

Hierarchies are commonly used to represent the clinical vocabulary [7]: for instance, ICD9 and COSTAR or the axes used in SNOMED. MeSH and HELP also use hierarchical representations to organize the clinical vocabulary, but they allow a number of terms to appear in more than one position. The pecularity of each hierarchy is influenced by its finality and its computational use. The consequent differences influence the relational organization of the terms: some hierarchies use only *is-a* relationships (ICD, CMIT), while others combine them with other relations, such as *is-a-parent*, *is-part-of*, *associated-with*, etc. in SNOMED [7].

Controlled clinical vocabularies often provide sufficient information and can be used as knowledge sources. Each concept represented in the controlled vocabulary structure is generally accompanied by its own semantics and definition. The concept specific location(s) within a structured clinical vocabulary and its relations with other concepts allow the improvement of the quality of the knowledge base. These features exist in several controlled clinical vocabularies, and in particular in UMLS. Several ways of using UMLS have been suggested. They permit to share experiences and to look for ways to enhance the use of UMLS [8, 7, 9]. Other UMLS users pointed out some limitations in using UMLS [5, 6]. These limitations concern specific clinical domain coverage, or specific use, *e.g.* patient record classification.

### Conceptual Graphs

The CG formalism is a knowledge representation language initially designed to capture the meaning of natural language. CGs have been used in many

natural language understanding works [10, 11, 12] as well as in specifically medical purposes [13]. CGs are used to build conceptual representations through networks where *concept nodes* represent entities, attributes, states and events linked by *conceptual relation nodes*. Every node has a label: a concept is noted within brackets, and a conceptual relation within parentheses, *e.g.*

$G1$ is `[artery:*]->(link)->[heart:#132]`

All the operations on CGs depend on the way labels can be compared.

We call *support* the possible labels for concepts and for conceptual relations. A concept label is made of a *type* and a *referent* whereas a conceptual relation label is a *relation*. There is a partial ordering over types corresponding to an *is-a* hierarchy. Types are organized in the concept type lattice (henceforth CTL). Referents are either individual markers, pointing to a given entity, or a variable (noted *), designing some entity. Therefore, a concept label subsumes another if its type and referent subsume or are equal to the type and referent of the other. The relations have no particular structure in the basic theory. In MENELAS, the relation hierarchical structure is a tree. The CG $G1$ mentioned above could be a representation for: *"There is some artery connected to the entity #132, which is a heart"*.

The subsumption relation over labels, mainly types, is the basis for CG operations. It induces a subsumption relation over CGs. In the following CGs: $G2$ is `[heart:*]`.

$G3$ is `[aorta:#45]->(link)->[heart:#132]`.

if we consider that *aorta* is a subtype of *artery* in the support, then $G2$ is a generalization of $G1$, and $G3$ a specialization of $G1$.

### Understanding Medical Language

Medical texts contain vast amounts of information. Their automatic analysis has been the subject of a large research area [14, 15]. Whereas most early methods were essentially syntax-based, recent approaches focus on the semantic representation of medical texts [11, 12, 16]. They enable a deeper level of understanding, including the representation of information that was implicit in the texts but is evident for the target readers [16].

The objective of MENELAS (European project AIM A2023) is to perform such an in-depth understanding of patient discharge summaries (PDSs) in French, English and Dutch, to allow users to access the information they contain. MENELAS adopts a knowledge-based approach to natural language understanding, and relies on a large body of medical knowledge to perform its task.

## THE STRUCTURATION OF KNOWLEDGE IN MENELAS

The MENELAS system relies upon the general hypothesis that understanding a PDS consists in building a conceptual representation of the world situation linguistically described in the text, by using conceptual knowledge. Such a hypothesis can be justified insofar as we are concerned with technical reports that describe what happens to the patient during a hospitalization. In this purpose, the natural language analysis subsystem of MENELAS uses both large coverage morpho-syntactic analyzers and elaborated semantic and "pragmatic" analyzers. The semantic analyzer deals with sentence semantics and handles the literal meaning of sentences. This analyzer builds a conceptual graph from a sentence by associating concepts to words thanks to a semantic lexicon. Roughly speaking, the semantic analyzer goes from words to concepts. The pragmatic analyzer relies on medical and common-sense knowledge to deal with implicit information: it uses pieces of knowledge to infer new concepts, representing additional information, from previous ones. Roughly speaking, the pragmatic analyzer goes from concepts to concepts.

As previously said, adopting the CG formalism constrains the way we represent knowledge in our system. While concepts have types that are hierarchically organized in an *is-a* lattice, relations are organized in an *is-a* tree. These structures are the backbone of knowledge representation in MENELAS and provide guidelines for acquiring this knowledge. We distinguish three basic types of knowledge: catalogs, schemas and scripts.

*Catalogs* contain pieces of knowledge which detect when a literal meaning found by the semantic analyzer has no pragmatic value: there exists no possible situation corresponding to what the graph represents. Such catalogs contain necessary truths, that must be satisfied by any representation. Any non-conforming graph is discarded. *Schemas* are pieces of knowledge that enable inference on the remaining graphs. Such inferences are intended to complete the conceptual representation. They can be either context independent or context dependent, *i.e.* by default. In the first case, these inferences are performed before the integration of what is being understood to the current text representation and in the second case, afterwards. Integration is guided by *scripts*, which capture knowledge about temporal aspects, *i.e.* the evolution of the situation. Since PDSs tell a story, temporal information is particularly important for integrating into a story representation the various pieces of information extracted from a text. Additional

inferences on the pathophysiological state of the patient are performed by a causal probabilistic network (CPN). Scripts and the CPN are not discussed here.

## BUILDING THE CTL

We describe here the construction of the CTL. The following section will show the way the CTL guides the acquisition of catalogs and schemas.

### The Contents of the CTL

The very philosophy of the CG formalism hinges on the concept type lattice and the relation tree, along with the representation in each concept of a type (intensional definition of the concept) and a referent (extensional definition of the concept). The overall organization of knowledge is determined in part by the CG formalism.

We found that the CTL building process must conform to several principles. First, the CTL captures context independent knowledge, that is, properties that are intrinsic to the concept being described. For example, the fact that the aorta is a kind of artery is always true, in any context. In contrast, that a disease like diabetes may be a risk factor for cardiovascular problems is not intrinsic to diabetes, but to the role it plays in the pathophysiological process under consideration. Second, *is-a* transitivity must obtain throughout the CTL. A common mistake consists in confusing *part-of* links with *is-a* links. Also, each new type should be introduced at the most specialized place where it is still a generalization of all its descendants. Third, the CTL contains only concepts but *no words:* each concept type should have a single meaning, even when it has multiple parents in the lattice; and each notion must be represented by a unique concept type. Linguistic phenomena like synonymy are handled by the semantic lexicon that attaches a same concept to various synonyms. The pragmatic analyzer and its knowledge are concerned only by conceptual representations and is independent of their linguistic realizations. The latter are handled by the semantic analyser.

### Using UMLS as a Resource

There exists no universally admitted method for building a CTL. Our approach is empiric [1], based on personal experience, on the experience of other teams [17, 11] and influenced by the task of our system. Validation is done manually on a case by case basis. We set as an objective to establish a well-structured higher part of the CTL, from which all other concept types will inherit, and an exhaustive lower part, consistent with both the lexicon and the knowledge base. The above principles guided the construction of the CTL.

Our starting point was the semantic hierarchy extracted from the UMLS semantic network, where only *is-a* relations have been kept [1]. The advantage was to build on an existing, large coverage categorization of medical concepts, established from numerous sources [4]. The other semantic relations contributed to our relation hierarchy.

### Adapting UMLS Locally

Categorization in UMLS is restricted to semantic types, contrasted with metathesaurus entries, and is thus moderately deep; we have taken some distance from this initial structure. On the one hand, our purpose is different from that of UMLS, and we had to cater for general concepts that are useful for natural language understanding. This adds structure to the higher part of the CTL. On the other hand, we focus on the domain of cardiology, and we had to enrich and refine some of the lower parts of the hierarchy. As a result, our CTL is deeper than and its structure is different from the UMLS semantic hierarchy.

Useful notions for natural language include *time, space, measurable entities, units of measure, attributes* and *values*. These entries could already be found in UMLS, except *measurable entities* and *units of measure*, but we inserted them directly under the root of the CTL. We have organized most other concept types in the general categories of *timed* (e.g., action, state) and *object* (e.g., organ, drug), according to whether or not they can be dated. However, below these, we have kept a large number of UMLS entries (and all the semantic types that they bear), but we reorganized them to fit our needs; the most noticeable ones are *anatomical structure, biologic physiologic and pathologic function, chemical viewed functionally, health care activity, finding*. We also removed several branches where concept types correspond to notions that are never mentioned in our PDSs (about 60 semantic types out of 134 in Meta 1.2). This is the case for *chemical* for which we only kept the functional part, for *organism*, for *activity*, for which we only kept *health care activity*. The exact number of concepts that we added is less meaningful, given the much finer granularity of our CTL. For instance, under *body part, organ or organ component*, we introduced *vessel* then *blood vessel* then *artery* then *coronary artery* then *circumflex artery*, which can be found in UMLS, but as metathesaurus entries. Our CTL currently contains 772 concept types.

We included the UMLS relations in our relation hierarchy, since they give a good description of medical semantics. We completed them with more general relations dealing with time, space, values and more general roles such as *agent* [10].

The above work relied on three sources of information: a PDS corpus, medical encyclopedias and dictionaries, and medical articles. We studied a *corpus* of one hundred PDSs coming from several hospitals, with two purposes in mind: (*i*) to get an exhaustive view of the notions used in the PDSs of the domain; these notions were abstracted from the most frequently used words and expressions; (*ii*) to know where to insert medical concepts in the CTL; we examined the context of use of these notions, in order to identify proximities (*is-a* relationships) between them. Starting from terms occurring in the PDSs, this step allowed to complete and sometimes reorganize the UMLS semantic tree. This method can be compared with [18], who starts from the metathesaurus concepts to build and refine the semantic network. The *medical encyclopedias* and *dictionaries* [19] help to check for internal CTL consistency and absence of redundancy. Finally, *publications* in medical journals help to understand the context of use of some notions, such as new therapeutic techniques.

## USING THE CTL FOR ACQUIRING KNOWLEDGE

The knowledge we use to understand a PDS does not consist of an unordered set of rules or unorganized atoms: this knowledge is highly structured by relations such as *is-a* and *part-of*, that define abstraction levels. In MENELAS, every piece of knowledge can be represented at the appropriate level of abstraction by being associated with a concept in the CTL, and inherited by the lower levels. We illustrate this with catalogs and schemas.

### Catalogs

In order to build the catalogs, we start from concepts in the higher part of the CTL and examine the relations they can have. Then, by manual specialization, using domain knowledge, we progressively go down to the lower parts of the CTL. Our starting point was once again a portion of UMLS: the set of constraints on the semantic types which can participate in a relation. This was possible since a large number of concepts in our CTL are borrowed from the UMLS semantic network. The principle here was to remain within the framework of constraints, which restrict the allowable conceptual representations but should not themselves bring any new information: catalogs filter, while schemas add information.

We illustrate the approach on concept type *fully_formed_anatomical_structure* and its descendants. A UMLS relation yields the constraint:

```
Cat. [fully_formed_anatomical_structure]-
        (location_of)->[pathologic_function]
```

By manual specialization in the CTL, we added

```
Cat. [artery]->(location_of)->[stenosis].
```

where *artery* is a specialization of *fully_formed_anatomical_structure*, and *stenosis* is a specialization of *pathologic_function* in the CTL.

### Schemas

Schemas are assembled from the concept types present in the CTL and correspond to notions occurring in the PDSs. They are possibly augmented by an expert as deemed necessary. By navigating through Meta-1, we obtained information on the contexts and definitions of terms. The two main principles in this part of the knowledge base were to avoid redundancy and to stop the description at the right level of detail. The guide here was again the information found in the PDS corpus. In the following simplified example schema,

```
Schema [stenosis:*x]<-(location_of)<-[artery]
is [stenosis:*x]-
      (attr)->[degree]-
         (val)->[number]->(unit)->[percent].
      (attr)->[length]
      (causes)<-[atherosclerosis].
```

the structure of the schema comes from the elements found in the PDSs for the first two relations. Additional pathophysiological knowledge was used to produce the last relation. The structure of this schema conforms to the level of detail found in PDSs; the pathophysiological precision on atherosclerosis allows the system to link a stenosis with other consequences of atherosclerosis, and in particular with clinical ones.

## DISCUSSION

The UMLS semantic hierarchy was the starting point of our CTL. UMLS's global view of medical language enabled us to begin with a large coverage, documented, shared basis. However, differences in purpose — integrating medical information resources vs understanding cardiology PDSs — have lead us to move towards a structure more adapted to our needs. We kept about half of UMLS's semantic types: the ones the most related with the patient (*anatomical structure, biologic physiologic and pathologic function, chemical viewed functionally, health care activity, finding* and their descendants), present in any health care domain. We reorganized them according to the needs of our application domain, in particular by creating concept types for time-related vs non-time-related notions. We also introduced in the lower parts of the CTL a level of granularity much finer than in the UMLS semantic hierarchy — although finer details can be found in Meta-1. The resulting CTL includes 772 concept types, with a depth of 12. These changes were motivated by the fact that the organization of our system gives the CTL a central role in the structuration and hence the acquisition of knowl-

edge. All the concept types and their hierarchical links are then used for the construction of the semantic lexicon and the knowledge base.

This work gave a central role to the *is-a* relation. More generally, one could think of a concept type lattice for other relations such as *part-of*. In fact, the crucial point is the *transitivity* of the relation, which enables inheritance and representational parsimony. The CG formalism relies upon the *is-a* relation because of the particular role of taxonomies for representing empirical knowledge.

One of the issues that we encountered involved notions such as risk factors, which, as pointed out above, are not defined intrinsically. Such a notion is in fact *relational*, and is best represented with a relation, here, *risk_factor_of*. We then use the schemas to define which concepts play the *role* of risk factors relative to which disease, by adding to their schemas relations such as the following:
`[diabetes]->(risk_factor_of)->[heart_disease]`
Finally, in the lexicon, the term "risk factor" will be defined as "something which bears relation *risk_factor_of* to a disease."

Having an important common kernel with UMLS, besides providing an initial coherent structure, allows the subsequent use of other UMLS components. It also facilitates the maintenance of the knowledge base as new versions of UMLS come out. The UMLS semantic relations other than *is-a* provide the starting point of our *catalogs*, and the Meta-1 terms definitions and contexts constitute as many clues for the structuration of the lower levels of the CTL and for the construction of *schemas*.

Reference


[1]. A. McCray. The UMLS semantic network. In SCAMC [20], pages 503–507.

[2]. A. L. Rector, W. A. Nowlan, S. Kay. Conceptual knowledge: the core of medical information systems. In Lun et al. [21], pages 1420–1426.

[3]. J. F. Sowa. Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley, London, 1984.

[4]. B. L. Humphrey, D. A. B. Lindberg. Building the Unified Medical Language System. In SCAMC [20], pages 475–480.

[5]. S. Huff, R. H. Warner. A comparison of Meta-1 and Help terms: Implication for clinical data. In SCAMC [22], pages 166–174.

[6]. C. G. Chute, Y. Yang, M. S. Tuttle, D. D. Sherertz, N. E. Olson, M. S. Erlbaum. A preliminary evaluation of the UMLS metathesaurus for patient record classification. In SCAMC [22], pages 161–165.

[7]. J. J. Cimino, G. Hripcsack, S. B. Johnson, P. D. Clayton. Designing an introspective, multipurpose, controlled medical vocabulary. In SCAMC [20], pages 513–518.

[8]. R. Appel, H. Komorowski, C. Barr, R. Greenes. Intelligent focusing in knowledge indexing and retrieval — the relatedness tool. In Proc of SCAMC'88, pages 152–157. IEEE, 1988.

[9]. J. J. Cimino, G. Hripcsak, S. B. Johnson, C. Friedman, D. J. Fink, P. D. Clayton. UMLS as a knowledge base — a rule based expert system approach to controlled medical vocabulary management. In SCAMC [22], pages 175–179.

[10]. J. Fargues, M.-C. Landau, A. Dugourd, L. Catach. Conceptual graphs for semantics and knowledge processing. IBM Journal of Research and Development, 30(1):70–79, 1986.

[11]. R. Baud, A. Rassinoux, J. Scherrer. Natural language processing and semantical representation of medical texts. Methods of Information in Medicine, 31:117–125, 1992.

[12]. M. Schröder. Knowledge-based processing of medical language: A language engineering approach. In Proceedings of GWAI'92, Bonn, D., September 1992.

[13]. K. Campbell, M. Musen. Representation of clinical data using SNOMED III and conceptual graphs. In SCAMC [23], pages 354–358.

[14]. N. Sager, C. Friedman, M. S. Lyman, editors. Medical Information Processing - Computer Management of Narrative Data. Addison Wesley, Reading, Mass., 1987.

[15]. J. R. Scherrer, R. A. Côté, S. H. Mandil, editors. Computerised Natural Medical Language Processing for Knowledge Engineering, Amsterdam, 1989. North-Holland.

[16]. M. Cavazza, L. Doré, P. Zweigenbaum. Model-based natural language understanding in medicine. In Lun et al. [21], pages 1356–1361.

[17]. J. Fargues, A. Perrin. Synthesizing a large concept hierarchy from french hyperonyms. In Proc COLING'90, pages 112–117, Helsinki, 1990.

[18]. A. McCray. Extending a natural language parser with UMLS knowledge. In SCAMC [23], pages 194–198.

[19]. M. Garnier, V. Delamare. Dictionnaire des termes de médecine. Maloine, Paris, 1992.

[20]. Proc of SCAMC'89. IEEE, 1989.

[21]. K. C. Lun, P. Degoulet, T. Piemme, O. Rienhoff, editors. Proc MEDINFO 92, Amsterdam, 1992. North Holland.

[22]. Proc of SCAMC'90. IEEE, 1990.

[23]. Proc of SCAMC'92. Mc Graw Hill, 1992.