


Building a Large Knowledge Base Semi-Automatically

Udo Hahn

 Text Knowledge Engineering Lab
Freiburg University
Werthmannplatz 1
D-79085 Freiburg, Germany
hahn@coling.uni-freiburg.de

Stefan Schulz

Medical Informatics Department
Freiburg University Hospital
Stefan-Meier-Str. 26
D-79104 Freiburg, Germany
stschulz@uni-freiburg.de

Abstract

We describe a knowledge engineering approach by which conceptual knowledge is extracted from an informal, semantically weak medical thesaurus (UMLS) and automatically converted into a formally sound description logics system. Our approach consists of four steps: concept definitions are automatically generated from the UMLS source, integrity checking of taxonomic and partonomic hierarchies is performed by the terminological classifier, cycles and inconsistencies are eliminated, and incremental refinement of the evolving knowledge base is performed by a domain expert. We report on experiments with a terminological knowledge base composed of 164,000 concepts and 76,000 relations.

Introduction

Over several decades, an enormous body of medical knowledge, e.g. disease taxonomies, medical procedures, anatomical terms etc., has been assembled in a wide variety of medical terminologies, thesauri and classification systems. The conceptual structuring of a domain they allow is typically restricted to the provision of broader/narrower terms, related terms or (quasi-)synonymous terms. This is most evident in the UMLS, the *Unified Medical Language System* (McCray & Nelson 1995), an umbrella system which covers more than 50 medical thesauri and classifications. Its metathesaurus component contains more than 600,000 concepts which are structured in hierarchies by 134 semantic types and 54 relations between semantic types. Their semantics is shallow and intuitive, which is due to the fact that their usage is primarily intended for humans engaged in various forms of clinical knowledge management.

Given its size, evolutionary diversity and inherent heterogeneity, there is no surprise that the lack of a formal foundation leads to inconsistencies, circular definitions, etc. (Cimino 1998). This may not cause utterly severe problems when humans are in the loop and its use is limited to disease encoding, accountancy or document retrieval tasks. However, anticipating its use for more knowledge-intensive applications such as natural language understanding of medical narratives (Hahn, Romacker, & Schulz 1999) or medical

decision support systems (Reggia & Tuhrim 1985), those shortcomings might lead to an impasse.

As a consequence, formal models for dealing with medical knowledge have been proposed, using representation mechanisms based on conceptual graphs, semantic networks or description logics (Volot *et al.* 1994; Mays *et al.* 1996; Rector *et al.* 1997). Not surprisingly, however, there is a price to be paid for more expressiveness and formal rigor, *viz.* increasing modeling efforts and, hence, increasing maintenance costs. Therefore, concrete systems making full use of this rigid approach, especially those which employ high-end knowledge representation languages are usually restricted to rather small subdomains.

The knowledge bases developed within the framework of the above-mentioned terminological systems have all been designed from scratch – without making systematic use of the large body of knowledge contained in those medical terminologies. An intriguing approach would be to join the massive *coverage* offered by informal medical terminologies with the high level of *expressiveness* supported by formal inferencing systems, as developed in the AI knowledge representation community, in order to develop formally solid medical knowledge bases on a larger scale. This idea has already been fostered by Pisanelli, Gangemi, & Steve (1998) who extracted knowledge from the UMLS semantic network as well as from parts of the metathesaurus and merged them with logic-based top-level ontologies from various sources. In a similar way, Spackman & Campbell (1998) describe how SNOMED (Côté 1993) evolves from a multi-axial coding system into a formally founded ontology. Unfortunately, the efforts made so far are entirely focused on generalization-based reasoning along *is-a* hierarchies and lack a reasonable coverage of partonomies.

Part-Whole Reasoning

As far as medical knowledge is concerned, two main hierarchy-building relationships can be identified, *viz.* *is-a* (taxonomic) and part-whole (partonomic) relations. Unlike generalization-based reasoning in concept taxonomies, no fully conclusive mechanism exists up to now for reasoning along part-whole hierarchies in description logic systems. For medical domains, however, the exclusion of part-whole reasoning is far from adequate. Anatomical knowledge, a central portion of medical knowledge, is principally organized along part-whole hierarchies. Hence, any proper

⁰Copyright © 2001, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

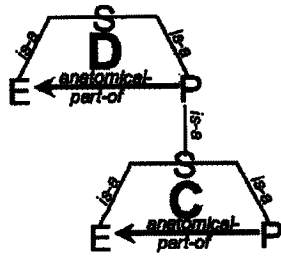


Figure 1: SEP Triplets: Partitive Relations within Taxonomies

medical knowledge representation has to take account of both hierarchy types (Haimowitz, Patil, & Szolovits 1988).

Various approaches to the reconstruction of part-whole reasoning within the object-centered representation paradigm are discussed by Artale *et al.* (1996). In the description logics community several language extensions have been proposed based on special constructors for part-whole reasoning (Rector *et al.* 1997; Horrocks & Sattler 1999), though at the cost of increasing computational complexity. Motivated by informal approaches sketched by Schmolze & Mark (1991) we formalized a model of part-whole reasoning (Hahn, Schulz, & Romacker 1999) that does not exceed the expressiveness of the well-understood, parsimonious concept language *ALC* (Schmidt-Schauß & Smolka 1991).¹

Our proposal is centered around a particular data structure, so-called *SEP triplets*, especially designed for part-whole reasoning (cf. Figure 1). They define a characteristic pattern of IS-A hierarchies which support the emulation of inferences typical of transitive PART-OF relations. In this formalism, the relation ANATOMICAL-PART-OF describes the partitive relation between physical parts of an organism.

A triplet consists, first of all, of a composite 'structure' concept, the so-called **S-node** (e.g., HAND-STRUCTURE). Each 'structure' concept subsumes both an anatomical *entity* and each of the anatomical *parts* of this entity. Unlike entities and their parts, 'structures' have no physical correlate in the real world — they constitute a representational artifact required for the formal reconstruction of systematic patterns of part-whole reasoning. The two direct subsumees of an S-node are the corresponding **E-node** ('entity') and **P-node** ('part'), e.g., HAND and HAND-PART, respectively. Unlike an S-node, these nodes refer to specific ontological objects. The E-node denotes the whole anatomical entity to be modeled, whereas the P-node is the common subsumer of any of the parts of the E-node. Hence, for every P-node there exists a corresponding E-node for the role ANATOMICAL-PART-OF. Some basic anatomical relations in terms of SEP triplets are illustrated in Figure 2.

¹*ALC* allows for the construction of hierarchies of concepts and relations, where \sqsubseteq denotes subsumption and \doteq definitional equivalence. Existential (\exists) and universal (\forall) quantification, negation (\neg), disjunction (\sqcup) and conjunction (\sqcap) are supported. Role filler constraints (e.g., typing by *C*) are linked to the relation name *R* by a dot, $\exists R.C$.

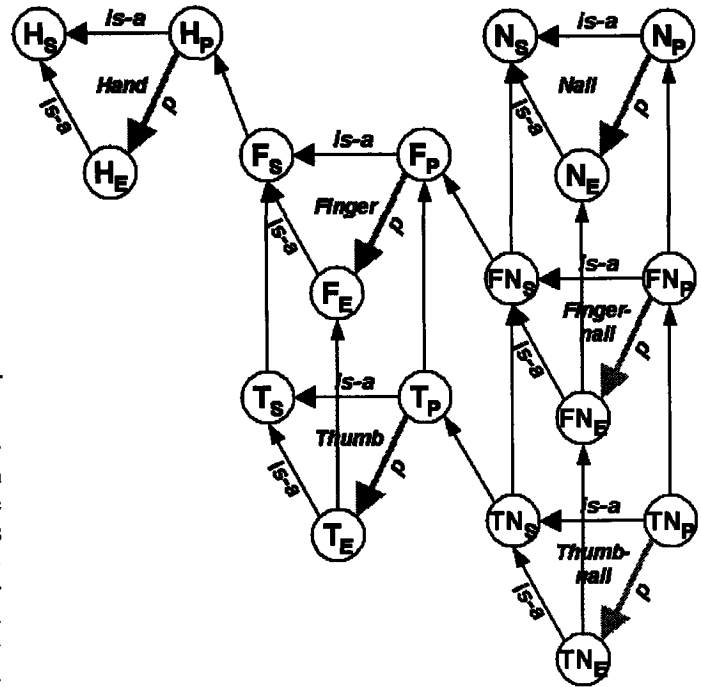


Figure 2: Partonomic Hierarchy of the Concept HAND

The reconstruction of the relation ANATOMICAL-PART-OF by taxonomic reasoning proceeds as follows. Let us assume that C_E and D_E denote E-nodes, C_S and D_S denote the S-nodes that subsume C_E and D_E , respectively, and C_P and D_P denote the P-nodes related to C_E and D_E , respectively, via the role ANATOMICAL-PART-OF (cf. Figure 1). These conventions can be captured by the following terminological expressions:

$$C_E \sqsubseteq C_S \sqsubseteq D_P \sqsubseteq D_S \quad (1)$$

$$D_E \sqsubseteq D_S \quad (2)$$

The P-node is defined as follows (note the disjointness between D_E and D_P):

$$D_P \doteq D_S \sqcap \neg D_E \sqcap \exists \text{anatomical-part-of}.D_E \quad (3)$$

Since C_E is subsumed by D_P (1) we infer that the relation ANATOMICAL-PART-OF holds between C_E and D_E :

$$C_E \sqsubseteq \exists \text{anatomical-part-of}.D_E \quad (4)$$

Knowledge Import and Refinement

Our goal is to extract conceptual knowledge from two highly relevant subdomains of the UMLS, *viz.* anatomy and pathology, in order to construct a formally sound knowledge base using a terminological knowledge representation language. This task will be divided into four steps: (1) the automated generation of terminological expressions, (2) their submission to a terminological classifier for consistency checking, (3) the manual restitution of formal consistency in case of inconsistencies, and, finally, (4) the manual rectification and refinement of the formal representation structures.

C0005847	CHD	C0014261	part of	MSH99	MSH99
C0005847	CHD	C0014261		CSP98	CSP98
C0005847	CHD	C0025962	isa	MSH99	MSH99
C0005847	CHD	C0026844	part of	MSH99	MSH99
C0005847	CHD	C0026844		CSP98	CSP98
C0005847	CHD	C0034052		SNM98	SNM98
C0005847	CHD	C0035330	isa	MSH99	MSH99
C0005847	CHD	C0042366	part of	MSH99	MSH99
C0005847	CHD	C0042367	part of	MSH99	MSH99
C0005847	CHD	C0042367		SNM2	SNM2
C0005847	CHD	C0042449	isa	MSH99	MSH99

Figure 3: Semantic Relations in the UMLS Metathesaurus

Step 1: Automated Generation of Terminological Expressions. Sources for concepts and relations were the UMLS semantic network and the *mrrel*, *mrcon* and *mrsty* tables of the 1999 release of the UMLS metathesaurus. The *mrrel* table which contains approximately 7,5 million records (cf. Figure 3) exhibits the semantic links between two UMLS CUIs (concept unique identifier),² the *mrcon* table contains the concept names and *mrsty* keeps the semantic type(s) assigned to each CUI. These tables, available as ASCII files, were imported into a Microsoft Access relational database and manipulated using SQL embedded in the VBA programming language. For each CUI in the *mrrel* subset its alphanumeric code was substituted by the English preferred term found in *mrcon*.

After a manual remodeling of the 135 top-level concepts and 247 relations of the UMLS semantic network, we extracted, from a total of 85,899 concepts, 38,059 anatomy and 50,087 pathology concepts from the metathesaurus. The criterion for the inclusion into one of these sets was the assignment to predefined semantic types. Also, 2,247 concepts were found to be included into both sets, anatomy and pathology. Since we wanted to keep the two subdomains strictly disjoint, we maintained these 2,247 concepts duplicated, and prefixed all concepts by ANA- or PAT- according to their respective subdomain. This can be justified by the observation that these hybrid concepts exhibit, indeed, multiple meanings. For instance, TUMOR has the meaning of a malignant disease on the one hand, and of an anatomical structure on the other hand.

As target structures for the anatomy domain we chose SEP triplets. These were expressed in the terminological language LOOM which we had previously extended by a special DEFTRIPLET macro (cf. Table 1 for an example). Only UMLS *part-of*, *has-part* and *is-a* relation attributes are considered for the construction of taxonomic and partonomic hierarchies. Hence, for each anatomy concept, one SEP triplet is created. The result is a mixed IS-A and PART-WHOLE hierarchy.

For the pathology domain, we treated *CHD* (child) and *RN* (narrower relation) from the UMLS as indicators of

²As a coding convention in UMLS, any two CUIs must be connected by at least a shallow relation (in Figure 3, CHiD relations in the column REL are assumed between CUIs). These shallow relations may be refined in the column REL_A, if a thesaurus is available which contains more precise information. Some CUIs are linked either by *part-of* or *is-a*. In any case, the source thesaurus for the relations and the CUIs involved is specified in the columns X and Y (e.g., MeSH 1999, SNOMED International 1998).

```
(deftriplet HEART
:is-primitive HOLLOW-VISCUS
:has-part (:p-and
FIBROUS-SKELETON-OF-HEART
WALL-OF-HEART
CAVITY-OF-HEART
CARDIAC-CHAMBER-NOS
LEFT-SIDE-OF-HEART
RIGHT-SIDE-OF-HEART
AORTIC-VALVE
PULMONARY-VALVE
```

Table 1: Generated Triplets in LOOM Format

taxonomic links. No part-whole relations were considered, since this category does not apply to the pathology domain. Furthermore, for all anatomy concepts contained in the definitional statements of pathology concepts the 'S-node' is the default concept to which they are linked, thus enabling the propagation of roles across the part-whole hierarchy.

In both subdomains, shallow relations, such as the extremely frequent sibling *SIB* relation, were included as comments into the code to give some heuristic guidance for the manual refinement phase.

Step 2: Submission to the LOOM Classifier. The import of UMLS anatomy concepts resulted in 38,059 DEFTRIPLET expressions for anatomical concepts and 50,087 DEFCONCEPT expressions for pathological concepts. Each DEFTRIPLET was expanded into three DEFCONCEPT (S-, E-, and P-nodes), and two DEFRELATION (ANATOMICAL-PART-OF-X, INV-ANATOMICAL-PART-OF-X) expressions, summing up to 114,177 concepts. This yielded (together with the concepts from the semantic network) a total of 240,764 definitory LOOM expressions.

From 38,059 anatomy triplets, 1219 DEFTRIPLET statements exhibited a :HAS-PART clause followed by a list of a variable number of triplets, containing more than one argument in 823 cases (average cardinality: 3.3). 4043 DEFTRIPLET statements contained a :PART-OF clause, only in 332 cases followed by more than one argument (average cardinality: 1.1). The resulting knowledge base was then submitted to the terminological classifier and checked for terminological cycles and coherence. In the anatomy subdomain, one terminological cycle and 2328 incoherent concepts were found, in the pathology subdomain 355 terminological cycles though not a single incoherent concept were determined (cf. Table 2).

Step 3: Manual Restitution of Consistency. The inconsistencies of the anatomy part of the knowledge base identified by the classifier could all be traced back to the simultaneous linkage of two triplets by both *is-a* and *part-of* links, an encoding that raises a conflict due to the disjointness required for corresponding P- and E-nodes. In most of these cases the affected parents belonged to a class of concepts that obviously cannot be appropriately modeled as SEP triplets, e.g., SUBDIVISION-OF-ASCENDING-AORTA, ORGAN-PART. The meaning of each of these concepts almost paraphrases that of a P-node, so that in these

	Anatomy	Pathology
Triplets	38,059	—
defconcept statements	114,177	50,087
cycles	1	355
inconsistencies	2,328	0

Table 2: Classification Results for the Concept Import

cases the violation of the SEP-internal disjointness condition was resolved by substituting the involved triplets with simple LOOM concepts, by matching them with already existing P-nodes or by disabling IS-A or PART-OF links.

In the pathology part of the knowledge base, we expected a large number of terminological cycles, as a consequence of interpreting the thesaurus-style *narrower term* and *child* relations through taxonomic subsumption (IS-A). Bearing in mind the size of the knowledge base, we consider 355 cycles a tolerable amount. Those cycles were primarily due to very similar concepts, e.g., ARTERIOSCLEROSIS vs. ATHEROSCLEROSIS, AMAUROSIS vs. BLINDNESS, and residual categories (“other”, “NOS” = *not otherwise specified*). These were directly inherited from the source terminologies and are notoriously difficult to interpret out of their definitional context, e.g., OTHER-MALIGNANT-NEOPLASM-OF-SKIN vs. MALIGNANT-NEOPLASM-OF-SKIN-NOS. The cycles were analyzed and a negative list which consisted of 630 concept pairs was manually derived. In a subsequent extraction cycle we incorporated this list in the automated construction of the LOOM concept definitions, and given these new constraints, a fully consistent knowledge base was generated.

Step 4: Manual Rectification and Refinement of the Knowledge Base. This step – when performed for the whole knowledge base – is time-consuming and requires broad and in-depth medical expertise. An analysis of random samples from both subdomains is currently being performed by a domain expert. The preliminary data we supply refer to the analysis of two random samples of each one-hundred anatomy and one-hundred pathology concepts.

From the experience we gained so far, the following workflow steps can be derived:

- *Checking the correctness of both the taxonomic and partitive hierarchies.* Taxonomic and partitive links are manually added or removed. Primitive subsumption is substituted by non-primitive subsumption whenever possible. This is a crucial point, because the automatically generated hierarchies contain only information about the parent concepts and necessary conditions. As an example, the automatically generated definition of DERMATITIS includes the information that it is an INFLAMMATION, and that the role HAS-LOCATION must be filled by the concept SKIN. An INFLAMMATION that HAS-LOCATION SKIN, however, cannot automatically be classified as DERMATITIS.

Results: Taxonomic links had to be removed from 8 out of 100 sampled pathology concept definitions, but from none of the anatomy concept definitions. On the contrary, in 68 cases from this sample, anatomy con-

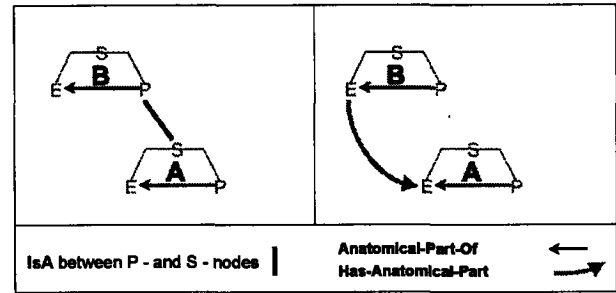


Figure 4: Part-whole Reasoning Patterns with SEP Triplets

cept definitions required the inclusion of anatomic or partonomic links. Often, necessary taxonomic or partonomic parents were already available, but not coded as UMLS parents or broader concepts. 7 from 100 anatomy concepts had to be considered as misclassified by the UMLS.

- *Check of the :has-part arguments assuming ‘real anatomy’.* In the UMLS sources *part-of* and *has-part* relations are considered as symmetric. According to our transformation rules, the attachment of a role HAS-ANATOMICAL-PART to an E-node B_E , with its range restricted to A_E implies the existence of a concept A for the definition of a concept B. On the other hand, the classification of A_E as being subsumed by the P-node B_P , the latter being defined via the role ANATOMICAL-PART-OF restricted to B_E , implies the existence of B_E given the existence of A_E . These constraints do not always conform to ‘real’ anatomy, i.e., anatomical concepts that may exhibit pathological modifications. Figure 4 (left) sketches a concept A that is necessarily ANATOMICAL-PART-OF a concept B, but whose existence is not required for the definition of B. This is typical of the results of surgical interventions, e.g., a large intestine without an appendix.

Results: The analysis of 15 triplet definitions that exhibit automatically generated :HAS-PART clauses revealed that 34% of the concepts should be eliminated from the :HAS-PART list in order not to obviate a coherent classification of pathologically modified anatomical objects.³ The opposite situation is also common (cf. Figure 4, right): the definition of A_E does not imply that the role ANATOMICAL-PART-OF be filled by B_E , but B_E does imply that the inverse role be filled by A_E . As an example, a LYMPH-NODE necessarily contains LYMPH-FOLLICLES, but there exist LYMPH-FOLLICLES that are not part of a LYMPH-NODE.

- *Analysis of the sibling relations and defining concepts as being disjoint.* In UMLS, *SIB* relates concepts that share the same parent in a taxonomic or partonomic hierarchy. Pairs of sibling concepts may have common

³In the example of Table 1, the concepts printed in *italics*, viz. AORTIC-VALVE and PULMONARY-VALVE should be eliminated from the :HAS-PART list, because they may be missing in certain cases as a result of congenital malformations, inflammatory processes or surgical interventions.

descendants or not. If not, they constitute the root of two disjoint subtrees. In a taxonomic hierarchy, this means that one concept implies the negation of the other (e.g., a benign tumor cannot be a malignant one, *et vice versa*). In a partitive hierarchy, this can be interpreted as *spatial disjointness*, viz. one concept does not spatially overlap with another one. As an example, ESOPHAGUS and DUODENUM are spatially disjoint, whereas STOMACH and DUODENUM are not (they share a common transition structure, called PYLORUS), such as all neighbor structures that have a surface or region in common. Spatial disjointness can be modeled such that the definition of the S-node of the concept *A* implies the negation of the S-node of the concept *B*.

Results: The large number of sibling concepts (on the average 7.3 siblings per concept in the anatomy, 8.8 in the pathology subdomain) makes the modeling of disjointness a time-consuming task, as every pair of concepts must be analyzed. At first glance, our data indicate that conceptual disjointness holds for at least two-thirds of the sibling concepts in both domains, and spatial disjointness for over three quarters in the anatomy domain.

- **Completion and modification of anatomy–pathology relations.** Surprisingly, only very few pathology concepts contained an explicit reference to a corresponding anatomy concept. These relations must, therefore, be added by a domain expert. In each case, the decision must be made whether the E-node or the S-node has to be addressed as the target concept for modification. In the first case, the propagation of roles across part-whole hierarchies is disabled, in the second case it is enabled.

Results: In an analysis of a random sample of 100 pathology concepts, only 17 were found to be linked with an anatomy concept. In 15 cases, the default linkage to the S-node was considered to be correct, in one case the linkage to the E-node was preferred. In another case, the linkage was considered false.

Conclusions

Instead of developing sophisticated medical knowledge bases from scratch, we here propose a ‘conservative’ approach — reuse existing large-scale resources, but refine the data from these resources so that advanced representational requirements imposed by more expressive knowledge representation languages are met.

The knowledge engineering approach we propose does exactly this. It provides a formally solid description logics framework with a modeling extension by SEP triplets so that both taxonomic and paronomic reasoning are supported equally well. While pure automatic conversion from semi-formal to formal environments causes problems of adequacy of the emerging representation structures, the refinement methodology we propose already inherits its power from the terminological reasoning framework. In our concrete work, we found the implications of using the terminological classifier, the inference engine which computes subsumption relations, of utmost importance and of out-

standing heuristic value. Hence, the knowledge refinement cycles are truly semi-automatic, fed by medical expertise on the side of the human knowledge engineer, but also driven by the reasoning system which makes explicit the consequences of (im)proper concept definitions.

Acknowledgements. We would like to thank our colleagues in the CLIF group for fruitful discussions and instant support. St. Schulz was partly supported by a grant from DFG (Ha 2097/5-2).

References

- Artale, A.; Franconi, E.; Guarino, N.; and Pazzi, L. 1996. Part-whole relations in object-centered systems: an overview. *Data & Knowledge Engineering* 20(3):347–383.
- Cimino, J. J. 1998. Auditing the Unified Medical Language System with semantic methods. *Journal of the American Medical Informatics Association* 5(1):41–45.
- Côté, R. 1993. *SNOMED International*. Northfield, IL: College of American Pathologists.
- Hahn, U.; Romacker, M.; and Schulz, S. 1999. How knowledge drives understanding: matching medical ontologies with the needs of medical language processing. *Artificial Intelligence in Medicine* 15(1):25–51.
- Hahn, U.; Schulz, S.; and Romacker, M. 1999. Part-whole reasoning: a case study in medical ontology engineering. *IEEE Intelligent Systems & their Applications* 14(5):59–67.
- Haimowitz, I. J.; Patil, R. S.; and Szolovits, P. 1988. Representing medical knowledge in a terminological language is difficult. In *Proceedings of the 12th Annual Symposium on Computer Applications in Medical Care – SCAMC’88*, 101–105.
- Horrocks, I., and Sattler, U. 1999. A description logic with transitive and inverse roles and role hierarchies. *Journal of Logic and Computation* 9(3):385–410.
- Mays, E.; Weida, R.; Dionne, R.; Laker, M.; White, B.; Liang, C.; and Oles, F. J. 1996. Scalable and expressive medical terminologies. In *Proceedings of the 1996 AMIA Annual Fall Symposium (formerly SCAMC) – AMIA’96*, 259–263.
- McCray, A. T., and Nelson, S. J. 1995. The representation of meaning in the UMLS. *Methods of Information in Medicine* 34(1/2):193–201.
- Pisanelli, D. M.; Gangemi, A.; and Steve, G. 1998. An ontological analysis of the UMLS metathesaurus. In *Proceedings of the 1998 AMIA Annual Fall Symposium – AMIA’98*, 810–814.
- Rector, A. L.; Bechhofer, S.; Goble, C. A.; Horrocks, I.; Nowlan, W. A.; and Solomon, W. D. 1997. The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine* 9:139–171.
- Reggia, J. A., and Tuhim, S., eds. 1985. *Computer-Assisted Medical Decision Making*, volume 1 & 2. New York: Springer.
- Schmidt-Schauß, M., and Smolka, G. 1991. Attributive concept descriptions with complements. *Artificial Intelligence* 48:1–26.
- Schmolze, J. G., and Mark, W. S. 1991. The NIKL experience. *Computational Intelligence* 6:48–69.
- Spackman, K. A., and Campbell, K. E. 1998. Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. In *Proceedings of the 1998 AMIA Annual Fall Symposium – AMIA’98*, 740–744.
- Volot, F.; Zweigenbaum, P.; Bachimont, B.; Ben Said, M.; Bouaud, J.; Fieschi, M.; and Boisvieux, J.-F. 1994. Structuration and acquisition of medical knowledge: using UMLS in the Conceptual Graph formalism. In *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care – SCAMC’93*, 710–714.