JAMIA
Reports from the Canon Group
Guest Editor: JAMES J. CIMINO, MD

Position Paper ■

Toward a Medical-concept Representation Language

DAVID A. EVANS, PHD, JAMES J. CIMINO, MD, WILLIAM R. HERSH, MD, STANLEY M. HUFF, MD, DOUGLAS S. BELL, MD, for the CANON GROUP

Abstract The Canon Group is an informal organization of medical informatics researchers who are working on the problem of developing a "deeper" representation formalism for use in exchanging data and developing applications. Individuals in the group represent experts in such areas as knowledge representation and computational linguistics, as well as in a variety of medical subdisciplines. All share the view that current mechanisms for the characterization of medical phenomena are either inadequate (limited or rigid) or idiosyncratic (useful for a specific application but incapable of being generalized or extended). The Group proposes to focus on the design of a general schema for medical-language representation including the specification of the *resources* and associated *procedures* required to map language (including standard terminologies) into representations that make all implicit relations "visible," reveal "hidden attributes," and generally resolve ambiguous or vague references. The Group is proceeding by examining large numbers of texts (records) in medical sub-domains to identify candidate "concepts" and by attempting to develop general rules and representations for elements such as attributes and values so that all concepts may be expressed uniformly.

■ **J Am Med Informatics Assoc.** 1994;1:207–217.

The purpose of this paper is to explicate the Canon Group's position and the principles that guide its work. The group name—"Canon"—reflects the Group's goal of establishing a basis for the canonical representation of medical concepts.*

Affiliations of the authors: Carnegie Mellon University, Pittsburgh, PA (DAE); Columbia Presbyterian Hospital, New York, NY (JJC); Oregon Health Sciences University, Portland, OR (WRH); University of Utah, Salt Lake City, UT (SMH); and Brigham and Women's Hospital, Boston, MA (DSB).

Correspondence and reprints: David A. Evans, PhD, Professor of Linguistics and Computer Science, Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213-3890.

Received for publication: 11/19/93; accepted for publication: 1/14/94.

*The Group uses "canon" in the sense of ". . . 4. a) an accepted principal or rule b) a criterion or standard of judgment c) a body of principles, rules, standards, or norms [LGk kan(o-)n, fr. Gk, model]" (Webster's 7th International Unabridged Dictionary).

As the title of this paper suggests, we believe that it is important for the medical community to have a common, uniform, and comprehensive approach to the representation of medical information—a "language" to use in expressing precisely the many details of medical observation, diagnosis, and patient management. We do not believe that the basis for such an approach exists today in any of the standard methods for the recording of medical data with controlled vocabularies (such as the Systematized Nomenclature of Medicine [SNOMED], the International Classification of Diseases, 9th ed. [ICD-9], etc.) or in extensions or combinations of standard vocabularies (such as may be possible with the Unified Medical Language System [UMLS]). While we strongly believe that an appropriate solution will support many automated-processing applications, including the direct processing of medical natural language, it is clear that the output of any particular automatic process, such as natural-language processing (NLP), will not in itself constitute a solution. Indeed, the quality of

the output of an automated process will depend on the ability of a system to represent and interpret medical concepts. Without a consensus on the notion of what a medical "concept" might be, and how it can be composed, and without an accepted standard for the representation of concepts in a form that can be manipulated by computers, there can be no automation. We therefore emphasize the need for a formal language—a "representation language"—with an associated set of meaning-bearing symbols (like "words"), a syntax (or rules that specify how words may be combined), and a semantics (or rules that make clear what relations words or their combinations have to one another). But we do not propose to create another "standard" vocabulary; the "medical-concept-representation language" (MCRL) we seek must be more like a programming language—a logical, well-defined system for manipulating concepts—than a human one. The essential work involves identifying and structuring medical concepts; a formal language is merely the most effective mechanism we can use to encode the "objects" that represent concepts for our purposes. We also believe that it will be possible to create such a language and to demonstrate its value in actual applications; but we recognize that the task is difficult and can be accomplished only over time: we must move toward our goal in successive steps of refinement and testing.

In the following section, we characterize the problem of concept representation as it manifests itself in practice, where we need comprehensive terminologies but do not find them and where we need applications that can apply to more than one task or more than one medical sub-domain but lack the "linguistic" bridges our systems require. In subsequent sections we consider, selectively, the limitations we find in controlled medical vocabularies, the need for an essentially semantic-conceptual basis for term representations, and the focus we have developed on an MCRL as the formal mechanism by which we hope to model concepts. We next present our methodology, including desiderata for our "science," and describe how we are proceeding individually and collectively to work toward what we hope will be a "merged model" for the representation of medical concepts. We conclude by comparing our effort with other contemporary projects touching on similar problems and by speculating about the possible outcomes of our work.

Ubiquity of the Problem

Our work has proceeded with a sense of urgency. The demands for information and record keeping in the health care system will only increase during this decade. To a large extent, the quality and effectiveness of health care will depend on the efficient processing and interpreting of medical language; the problem is ubiquitous. Consider, for example, the following scenarios:

1. A hospital has installed new decision-support tools. The hospital information system will check for contraindications to new orders entered into the system. One such target alert may be a warning to the ordering clinician whenever a nonsteroidal anti-inflammatory agent is ordered for a patient diagnosed as having peptic ulcer disease. The system designer must ensure that the system will recognize every existing nonsteroidal anti-inflammatory drug by name (and keep track of the steady stream of newly introduced ones) as well as every possible reference to acid-peptic disease.

2. A literature-retrieval system is built that will attempt to recognize concepts that are synonyms. A query is entered about calcium-channel blockers and their use in stroke. The system designer discovers that writers of articles use many synonymous terms for calcium-channel blockers, such as "calcium blockers," "calcium antagonists," and the individual names of the different agents. Likewise, stroke has many synonyms, such as "cerebrovascular accident" or "CVA," and may be referred to generally as "cerebrovascular disease."

3. A group is responsible for outcomes research. It is desirable to track all patient data, including symptoms, yet not have to process charts manually. A computer program is designed to extract information from patient records, but the system is ineffective because of variation in the descriptions of patients' symptoms. For example, what is written as "postprandial stomach pain" in one chart is described as "abdominal pain after meals" in another.

4. A renal dialysis center wants to develop a medical record that will support observational and controlled trials as a part of routine patient care. Much of the patient information will be collected by nurses and physicians through structured data entry to ensure that study parameters are rigorously assessed. The designers require a standard source of possible concepts, symptoms, and corresponding values in order to integrate different trials with overlapping data elements and to share their data with other participating centers.

The four scenarios reflect four different application areas in medical informatics—alerts, information retrieval, outcomes research, and predictive data en-

try—yet a principle obstacle to their implementation is the absence of consistent and comprehensive *language*. In particular, normal variation in the use of language, ambiguity, vagueness, and ellipsis—all of which people resolve easily when speaking or reading by appealing to *context*, the surrounding information that clarifies how a word or phrase might be interpreted—have significantly impeded wider use of computer applications in medicine. However, existing clinical vocabularies and classifications have, at best, limited notions of domain context or relationships between medical concepts and pertinent modifiers.

Despite the successful demonstration of prototypes in the areas of expert systems, automated information-retrieval systems, and patient databases, very few large-scale systems have been created or are being used. One reason, in virtually every case, is the difficulty presented by language representation when changing the scale (as in moving from a "toy" world to a hospital-sized setting) or the application (as in using a system designed to perform diagnosis as an aid to literature searches).† This is so because system designers often represent "only what is necessary" for their particular domains,[2] and it is much easier to create language-processing/representing components that cover the needs of the specific application than to design such components based on general principles or for universal application. Unfortunately, scalability and cross-application adaptability may have different requirements on data structures, so the solution of one problem does not necessarily lead to the solution of the other. It is time, we believe, to focus on the design of a general schema for medical-language representation and to develop the resources required to support changes in scale and translations of applications.

Inadequacy of Current Vocabulary Models

There are many examples of attempts to produce comprehensive controlled vocabularies, such as the International Classification of Diseases, 9th ed., with Clinical Modifications (ICD9-CM),[3] SNOMED-III,[4] the Gabrieli Nomenclature,[5] and the Read Clinical Classification.[6] While they have often succeeded in the tasks for which they were intended, none is appropriate for the range of tasks of interest to the Canon Group and all can be criticized for shortcomings.[7,8] In particular, individual efforts by Canon Group members (and other medical informatics professionals in our community) to employ such vocabularies in computational applications have met with limited success. The content and structure of existing vocabularies are certainly problematic,[9] but more fundamental issues are involved. For example, the practical problem of vocabulary maintenance and extension is rarely addressed. Given a term that is not included in the vocabulary, how is a new term to be accommodated? Should it be listed as a synonym of an existing term, or is there a subtle difference that should be preserved? Where the vocabulary begins to enumerate variations of a more general phenomenon (such as types of fractures in specific parts of specific bones), should new variations be represented through extension of the vocabulary or through some arrangements of existing terms? If the latter, what rules exist for doing so? What should be done when the vocabulary's granularity fails to match the application at hand?

Existing approaches address some such problems. For example, SNOMED[4] separates terms into "atomic" units. The units are organized into chapters or "axes," which can be combined to form complex concepts. Thus, SNOMED terms are compositionally extensible, but the rules for the process are imprecise. For example, in SNOMED one can "legally" represent "acute appendicitis" in any of the following forms:

1.	D5-46210	01	*Acute appendicitis, NOS*
2.	D5-46100	01	*Appendicitis, NOS*
	G-A231	01	*Acute*
3.	M-41000	01	*Acute inflammation, NOS*
	G-C006	01	*In*
	T-59200	01	*Appendix, NOS*
4.	G-A231	01	*Acute*
	M-40000	01	*Inflammation, NOS*
	G-C006	01	*In*
	T-59200	01	*Appendix, NOS*

Thus, we observe, while the efforts at cataloging *vocabulary* are impressive in SNOMED, the implicit *model* of what constitutes a valid medical concept is incomplete.

The examples from SNOMED illustrate another problem that is difficult to manage in most controlled vocabularies—the different roles that the same superficial expression may play in different contexts. When we record a *diagnosis*, the terms we use have a very special epistemological status: we implicitly assert the fact of the condition and invite all medically valid inferences. If we write "the patient had acute appendicitis" in a chart as part of a diagnostic sum-

†A good example of how language variation impedes performance when scaling up can be found in the article by Blair and Maron,[1] which attributes the poor results of an information-retrieval system in a moderate-scale application in large measure to the human tendency to talk or write or formulate queries about the "same" information using "different" words to express ideas.

mary, we may certainly be expected to treat the patient as though he or she had acute appendicitis. But, if we write "the patient was evaluated for acute appendicitis," we describe a clinical indication, which is distinct from asserting a diagnosis, especially as that diagnosis may have been ruled out under evaluation.

The distinctions in the examples above (1–4) derive from the difference between an *observation* of a process (e.g., inflammation) in a location (e.g., the appendix) and alternative ways for naming a diagnosis. Some forms of expression (1 or 2) are more accurate epistemologically when we are assessing a patient; SNOMED provides terms in axis "D," for diagnoses. The other forms (3 or 4) are more accurate when we are making observations; SNOMED provides terms in axes "M" and "T," for descriptions of morphology in "locations." In fact, while SNOMED does not give us the information we need in order to know when one form (and not another) should be used, we applaud the fact that it accommodates such distinctions; they reflect the essential science of clinical medicine[10,11] and are absolutely essential to the model of language use—the model of information—that we are interested in capturing in our work.

Concept Representations versus Terms

The problems we see in current and traditional controlled vocabularies derive from a failure to distinguish surface-form (nominal) expressions from the concepts that these expressions sometimes designate. In order to account for surface-form medical expressions, we must first understand and specify the structure of medical concepts. Such an undertaking is naturally challenging.

It is certainly difficult to talk about concepts (though everyone seems to have intuitions about what they are); it is even more difficult to establish a formal definition that all will accept. Nevertheless, we feel an obligation to state what we mean by *concept*; in the process it should be apparent that we require much more of a formal apparatus than a list of terms can provide to capture the "concepts" of medicine.

Let us begin by considering a concrete example. One can define a particular concept, such as *coarctation*, by referring to other concepts, such as *narrowing* or *constricting* and *blood vessel*. In this case, the other concepts must bear a specific relation to one another for the target concept to be realized: the constricting must be in (or of) the blood vessel for the concept *coarctation* to be instantiated. We see that, to understand or represent a concept such as *coarctation* properly, we need to evaluate objects, attributes, values, and relations. In fact, it is in the identification and evaluation of the objects, attributes, values, and relations that we discover (or not) that the concept applies.

Church[12] provides a useful characterization: a concept is, effectively, a decision procedure. Any object or set of objects may have a distinctive identity. Such a distinguished object or set of objects, in turn, may be associated with a symbol, which may be a *name*, such as "coarctation." In addition, there is a rule that can be used to determine whether anything we may encounter is a member of the set associated with the symbol. The set is the denotation of the symbol (possibly a name) and the rule that determines membership is the concept associated with the symbol (or name).

Notice that the ability actually to associate a name with a concept is incidental to the identity of the concept. Many concepts (e.g., mathematical ideas or images, such as particular faces) have no names. What is important is the procedure that determines whether something (a particular object) satisfies the conditions of membership in the set denoted by the concept. Clearly, to stipulate a procedure, one must first enumerate the elements that can be considered in identifying objects and in establishing membership in a set. As we can see with *coarctation*, complex attributes such as *narrowness* must be present in very specific relations to an anatomic object, *blood vessel*, for a particular phenomenon to be admitted to the set denoted by "coarctation."

With most standard medical vocabularies, the notion of a concept is implicit; the vocabularies have been concerned principally with named concepts that distinguish medicine from other disciplines. For the most part, the elements that are required to evaluate concepts are not represented. Since one of our principal goals is the automation of the processing of medical information and since automated processing will require evaluation of relations and the inferences they can support, we cannot use standard vocabularies in their present forms to model the concepts of medicine. Thus, we are less concerned with the identification of an actual vocabulary or set of terms than with the formal representations we must use to express precisely and explicitly the many semantic objects we find in the domain of medicine.

Computational Modeling of Medical Concepts

The representation of concepts, of course, has been a principal concern of researchers in artificial intel-

ligence for some time. (See, for example, the papers in Findler.[13]) Work on this problem has focused on a number of issues, such as the distinctions that must be made between the representation of terms and the uses of such representations (e.g., Brachman et al.[14]) and the need to ensure computational tractability in semantic networks.[15,16] A common approach involves the splitting of knowledge representations into a less-expressive terminological component where tractability can be assured and a more general assertional component that is used only judiciously. (For an overview, see MacGregor.[17])

When confronting the practical problem of building computational systems to manipulate concepts, it is typically difficult to realize the idealized view of concepts (as presented in the previous section). While it has been shown, for example, that a purely terminological representation is not adequate for a majority of the needs in medical decision support,[2] many computational systems still rely on term sets to substitute for concepts. Let us consider, briefly, how several groups have attempted to solve the problem of managing more complex representations of concepts in computationally tractable applications.

Among the first attempts to apply computational linguistic techniques to the mapping of free texts (in hospital charts) to database-entry representations was the work of Sager and coworkers in the Linguistic String Project.[18] The project represented the structure of medical information by developing frames that specified common predefined semantic relations among the types of concepts in the domain. The representations were developed, in part, through linguistic analysis of actual texts. While the approach did not specify a separate conceptual level, it underscored the importance (and power) of lexical decomposition of expressions and the need to capture semantic relations among concepts.

One of the first efforts to abstract away from specific terms to term representations for use in modeling general medical language was developed by Evans in the MedSORT project.[19,20] In MedSORT work, a distinction was made between the *lexical information* that is necessary to write and recognize words and phrases (such as those found in medical records), the *concepts* (which are abstract) that lexical items may combine to form, and the *contexts* in which concepts might appear (e.g., as a clinical observation or a description of a disease). To represent concepts, MedSORT required both an *ontology* or semantic classification scheme and rules that determined how elements in the classification scheme might combine. The link between lexical items (or vocabulary) and concepts was made possible by requiring every lexical item to have a *semantic type* and every concept representation to be expressed in terms of constellations of semantic types. Such constellations—in the form of semantic frames—made it possible to express precise relations among the elements of concepts and also to express relations among complex concepts and to distinguish similar concepts used in different contexts. In effect, the approach in MedSORT was to define an explicit *formal language* for medical concepts, encompassing a *grammar* for concept formation and structural or contextual constraints on *interpretation*.

More recently, Rector and coworkers in the GALEN project[21–23] have analyzed the problem as having three levels or domains:

1. The *terminology (vocabulary) model* proper.

2. The *conceptual model* that uses that terminology (vocabulary) to represent information.

3. The *knowledge base* that supports reasoning about events and objects in the real world.

The terminology model records the linguistic characteristics of medical concepts such as synonyms, homonyms, and parts of speech. In effect, it provides the domains for the *fields* or *entities* of the conceptual model. The conceptual model represents the set of rules or procedures for how the elements of the terminology can be combined to create valid expressions of medical information. The conceptual model corresponds to a grammar of medical concept representation or a logical (but not physical) database design. The knowledge base contains information, such as the normative characteristics of diagnoses, the relations of diseases to one another, and other rules or knowledge necessary to create abstractions from or use information represented in a database. The knowledge base expresses its knowledge by reference to concepts in the conceptual model.

In a similar vein, Friedman, Cimino, and coworkers have analyzed the problem of clinical vocabulary management as requiring distinctions among three levels[24,25]:

1. The *conceptual level*, which is the conceptual model, representing the contextual information, the structure of the concepts, and the naming of the concepts and synonyms and specifying the semantic typology and relations among concepts.

2. The *linguistic level*, associated with the words and phrases used to express the concepts, consisting of (a) a semantic typology for the words, (b) the rules

that enumerate the possible relationships among the semantic types, and (c) the rules that specify how complex phrases may be composed.

3. The *encoding level*, which specifies how linguistic expressions map to concepts.

All the efforts noted here have encountered the need to manage variation in language and to remain flexible in the identification of concepts, while addressing practical goals of producing data structures (representations) that may be used in computationally diverse applications.

Focus on an MCRL

Both MedSORT and GALEN have emphasized what we take to be the general goal of our efforts: the specification of the *resources* and associated *procedures* required to map language (including standard terminologies) into representations that make implicit relations "visible," reveal "hidden attributes," and generally resolve ambiguous or vague references. We aim to specify the means by which we may identify and characterize in a formal and uniform representation all valid, and only all valid, medical concepts in any string of text, whether natural language or controlled vocabulary. The resources we require to achieve such functionality are minimally:

1. A list of the basic lexical (atomic) units—the words, atomic phrases, abbreviations, symbols, etc.—in the medical domain.

2. A list of basic conceptual units—disambiguated target concepts—to which lexical units map.

3. A typology of basic concepts—providing, minimally, semantic types for each concept.

4. A network of general medical concepts, in which implicit relations between elements are made explicit.

The procedures we require are minimally:

1. Rules for the assignment of lexical items to specific conceptual units.

2. Rules for the composition of concepts—specifying how basic units come together to form more complex concepts.

Together, the resources and procedures specify a grammar (in a formal linguistic sense)—the grammar for the formation of medical concepts.

As an illustration, we note that to accommodate a phrase such as "pulse 45," a medical vocabulary system will need to identify "pulse" and "45" as lexical elements; to associate "pulse" with the measurement of a physiological phenomenon, *pulse-rate*, and "45" with a numeric designation, *numeric-count*, to express the "hidden" interval of the measurement (a minute); and to show the relations among all the representational parts—capturing, in effect, that the medical observation is of a pulse rate of 45 "pulses" per minute. In contrast, in representing the deceptively similar phrase "pulse strong," the system will need to identify "pulse" and "strong" as lexical elements; to associate "pulse" with the measurement of a physiologic phenomenon, *pulse-pressure*, and "strong" with a relative measurement, *relative-degree*, on a scale of strength, to express the "hidden" interval of the measurement (a beat), and to show the relations among elements—capturing the fact that the observation is of pulse pressure strength, measured as "strong" over each beat.

Given the Canon goal of cross-application design, it is essential that a Canon MCRL accommodate any valid medically meaningful expression, whether it has been "recorded" previously or not and whether it is expressed in a controlled vocabulary or in natural language. This is the requirement, effectively, of a medical *interlingua*. Others have explored the design of such representations in medical applications.[26–28] To serve as an interlingua, a representation must account for elements at the level of lexical items and must provide sufficient semantic structure (expressed in terms of concepts and their relations) to "cover" all of the required relations in language sources.

Striking the balance between minimally sufficient representations and the recording of all medical knowledge is a principal challenge in our work. There are numerous medical facts that have no direct bearing on the representation of medical language. However, many of the most important (and difficult) knowledge representation issues in medicine do reflect phenomena that must be accommodated in an MCRL. For example, models of time, causality, anatomy and physiology, spatial relationships, and uncertainty are directly and often subtly invoked in the use of modifiers or in the specification of the values of attributes. (The previous examples of "pulse 45" and "pulse strong" provide only the most superficial glimpse of such deeper problems.) No existing lexicon has a consistent set of modifiers for representing such implicit attributes of medical concepts; no knowledge-representation effort has succeeded in modeling such phenomena coherently or comprehensively. One facet of our task will be to explicate the relations between the ontology of biomedicine

and the concept models we require for language representation.

Desiderata for a Canon Methodology

It is historically natural, perhaps, that most medical terminologies have resulted from small group efforts. In such cases, consensus is valuable; objective scientific validity may be difficult to establish and ultimately unnecessary for the pragmatic task at hand. In the Canon effort, however, with its dependence on widespread collaboration and with its inherently abstract requirements (where intuition may be a poor guide to the "correctness" of a representational formalism), *how* we develop representations may be more important than *what representations* we produce. In particular, we agree on the need for:

1. A scientific methodology. We need to establish procedures, define methods and tools, and conduct our work so that our results are (a) reproducible, (b) extensible, (c) testable, (d) expressive, and (e) understandable.

2. An empirical basis. We need to work with (a) real data in (b) statistically meaningful samples that are (c) representative of sub-domains (implying depth) as well as (d) comprehensive for the general field of medicine.

We recognize the need to examine perhaps 100 subdomains of medicine with samples of 10 to 100 mb of text for each sub-domain. The empirical data will be our touchstone.

3. Completeness and coherence of coverage. We need to develop (a) a typology of concepts in medicine with sufficient (b) granularity and (c) breadth to ensure that all medical concepts can be represented.

In addition to naturally occurring medical texts, we aim to use existing nomenclatures and vocabularies as source corpora. This will be one means of ensuring that the coverage in our representation schema will be at least equivalent to what is currently available. Our schema will minimally enhance current terminologies by elucidating the tacit knowledge that has been compiled into them.

4. An interdisciplinary perspective. We need to bring resources and techniques from computer science, computational linguistics, and cognitive science, as well as medicine, to bear on the problem.

5. A focus on realistic outcomes. We need to keep an eye on the practical (and large-scale) applications that will use the resources we develop.

The test of our work will be the applications it can support. The most interesting long-term goal may not be the most realistic intermediate-term outcome. Even shorter-term results may be able to enhance performance in patient-record systems. We should be sure our efforts and methods are focused by attainable and desirable goals that contribute (perhaps stepwise) to longer-term progress.

6. Representational simplicity. We need to avoid adding more structure to the representations we develop than the minimum required for the task at hand.

The design of something as simple as a lexicon or as complex as a semantic network can easily become an end in itself. Once structure is in place, it may take on a life of its own. Worse, structure always reflects a theory of the relations among objects, which may be difficult to state, obscure to the outside observer, and epistemologically problematic. (Partial truth is harder to control—and eradicate—than falsehood.) Much that we hope to do practically may be possible with appropriately chosen minimal enhancements to terminologies. We plan to exploit such enhancements first.

The realization of our goals will require both a sustained effort and a group strategy. We feel that our work should proceed both from the bottom up (gathering empirical data and classifying them to establish a lexicon and corpus of phrases) and from the top down (characterizing domains, organization principles, and perspectives to establish appropriate classification schemes). The results of the former effort will provide the actual content material that will allow large-scale development of informatics applications, while the results of the latter work will provide the organizational structure for the content material.

We have no religious commitment to a particular representational formalism; we recognize the essential equivalence of semantic networks, frame systems, conceptual graphs, and logical-statement languages. But, we see the need for some form of data representation to express the complexity of concept relations precisely and to permit the sharing of our work across sites.

How We Are Proceeding

The discussions that led to the formation of the Canon Group occurred in late 1991. Group participants included members of several different laboratories, working in a variety of application areas. Despite differences in background and focus, it seemed to all that collaboration would be fruitful.

Over a period of two years, members of the Group have exchanged more than 1,000 electronic messages. The group has held several workshops. Topics have been explored that reflected individual interests, such as NLP, anticipatory data entry, information retrieval, knowledge representation, and decision support—all with their special requirements for language interpretation. It has become clear that similar language representation problems have to be solved in different applications and that the lessons learned in one application might apply to others.

Canon Group members have conducted pilot experiments to test Canon approaches and to exercise Canon principles.‡ In our first efforts, we have focused on the modeling of concepts in radiology; the selection of this domain reflects convenience (and available data) and the special interests of several members of the group. We have attempted to explore the domain both to develop an understanding of its complexity and to establish procedures that we may be able to exploit in subsequent work.

In the specific case of radiology, we collected a total of approximately 16 mb of radiology reports from four geographically distributed sites.§ We used a natural-language-processing system (CLARIT[29]) with the ability to produce thesauri automatically[30] to identify words and phrases that contained "medical information." The analysis was exhaustive; more than 55,000 unique candidate "terms" were discovered, of which more than 21,000 could be nominated for a composite thesaurus. The collection of expressions was more complete than one sees in standard vocabularies. For example, there were 320 unique terms containing the phrase "pleural effusion" and 211 unique terms containing the word "pneumonothorax." The use of such a large volume of source data in a sub-domain of medicine made it possible for us to satisfy the desiderata of empirical basis and completeness and coherence of coverage.

Some members of the Group have started to use the discovered terms to develop a semantic network for radiology. Others have used the terms to suggest semantic types for concepts and to validate proposals for "grammars" of semantic type combination. All have examined a small number of reports and attempted to (1) identify concepts and (2) express in formal notation all the relevant explicit and implicit relations that exist among the concepts. This has exercised the desideratum of representational simplicity.

Our first experiences have suggested a general methodology. As a group, we are now exploring several additional sub-domains of medicine and several application areas to help us understand the requirements of our task. In a top-down mode, we attempt to define general domains for which terminologic work is needed and to model those domains with an eye to the types of information we will need for our respective applications. We deliberately focus on domains that are of interest in multiple application areas. In a bottom-up mode, we choose a domain and obtain actual medical data for that domain (e.g., coded vocabularies or free text) from our own laboratories or in public repositories. Individually, we examine data with two goals in mind:

1. To determine where the top-down model needs expansion to deal with the complexities that occur as we attempt to manipulate real data with real applications, and

2. To develop content (e.g., lexical information, concepts, relations) that is needed to process the data.

When individual modeling has been completed, each investigator presents his or her model—encompassing lexical forms, conceptual structures, and framing contexts—to the group. Where overlap occurs, we are trying to merge our individual models. Where distinctions occur, we are attempting to expand our evolving "merged" model to embrace them or provide a means for establishing transformations between them. For example, one application may treat "appendicitis" as an atomic concept, while another may require decomposition into finer-grained elements, such as might be represented schematically as

"[Inflammation [Acute]]—[In]—[Appendix]"

We recognize that to relate one to the other we need a general transformation, such as

[*disease concept*]⇒
[*morphology concept*—`modified by`—*temporal concept*]
—`location relation`—
[*anatomy concept*]

Such a transformation allows us to recognize the different perspectives of different applications, while enabling us to make explicit the underlying conceptual details required if we are to share terminology

‡The tasks and some of our results are reported in this issue, for example.

§The sources of radiology reports were Brigham & Women's Hospital, Columbia Presbyterian Hospital, Latter Day Saints' Hospital, and Oregon Health Sciences University Hospital.

and exchange our interpreted patient data. Several aspects of this approach are worth noting.

First, while the preceding example appears to involve the same concepts that are given in the SNOMED examples presented previously, it differs in providing an explicit function for the transformation of the superficial vocabulary form into a generalized and precise representation. It might be noted that SNOMED terms could be used to "fill in" the disease, morphology, temporal, location, and anatomy constituents in the transformation frame. SNOMED clearly provides a rich source of medical concepts. Individually, the two applications of our example might well use SNOMED to satisfy their individual language-coding requirements. However, without the transformation function, we would not be able to share data across applications. To establish general transformation functions, in turn, we need deeper knowledge representation. For example, to apply such transformations, we need to know which temporal concepts can be applied to which classes of morphology.

Second, it is important to note that each individual investigator's model is developed using a common collection of documents and a common set of examples from the collection. In this way, we are able to understand the individual representations created by others. Obviously, we all share a model at the "human" level for the data we have explored, for we can easily discuss them. Our applications, however, reflect different simplifications away from our common, human understanding; this frustrates the creation of a simple shared computational model. By using a common data set, we enable our discussions to transcend the limitations imposed by our individual, computer-based approaches. Consider the difference between this approach and one in which we might assign five groups the task of providing a list of the 1,000 most common diseases. The result of such a task would be five lists encompassing 5,000 terms, some identical and synonymous, some apparently identical and non-synonymous, and others apparently different but actually synonymous. Resolving the five lists into a single list would be extremely difficult.

Third, the use of actual patient data, while often complex and sometimes messy, reassures us that our efforts are worthwhile. Using real data forces us to address "real-world" problems that might be overlooked if "toy" problems were chosen instead. Also, if we can model real data, then we will necessarily focus on the practical aspects of the task. Rather than becoming hopelessly mired in attempting to model everything that *might* be encountered, we are able to develop our models for those concepts that are *commonly* encountered, thus creating resources that can be immediately useful in our systems.

Distinction in Efforts

Some readers may view the Canon effort as potentially duplicating the efforts planned for the newly formed Computer-Based Patient Record Institute (CPRI). However, since the CPRI is initially interested in identifying existing technological solutions that can be applied to (among other things) the development of a coding scheme for patient records, the two efforts may prove to be complementary. The ability to model such a coding scheme is a stated goal of the Canon Group; some of its members are active participants in the CPRI effort. We expect our work to contribute to, rather than compete with, the CPRI's efforts.

Other readers will associate our work with the Unified Medical Language System (UMLS) project of the National Library of Medicine (NLM). There are indeed many similarities, and several Canon members are involved with developing or evaluating the UMLS; however, there are clear differences in the goals of the two efforts.

The UMLS is an attempt to develop a broad-based method for the integration of existing controlled medical vocabularies to facilitate access to and transfer between computer-based information sources.[31] The model includes the notion of concepts (which are specific meanings), terms (which are various synonymous names for the concepts drawn from disparate vocabularies), and strings (which are various lexical variants for particular terms found within the vocabularies). It also includes the notion of semantic typing for the concepts and the identification of possible semantic relations among semantic types. Terms from disparate vocabularies are related to each other either by being synonyms of the same concept or by mapping to separate UMLS concepts that are themselves related. The inter-concept relations are "broader-narrower" and "other." Specific semantic relations among particular concepts are present but not prevalent. Since its inception, the UMLS has continued to grow to incorporate additional terms from additional vocabularies, while the model has remained essentially stable (although the actual implementations of the model have evolved).

Some of the representational needs of our work are not presently among the priorities of the UMLS project. These include:

- A broad lexicon for NLP that extends beyond the scope of the UMLS source vocabularies.
- A mechanism for decomposing complex concepts lexically into their component concepts.
- A conceptual model for allowing the representation of new concepts without having to simply enumerate new permutations, as is done in the source vocabularies in the UMLS.
- An extensive "fleshing out" of semantic relationships at the inter-concept level, not just at the semantic class level.

The UMLS is clearly moving to address some of the above issues, for example, by including the specialist lexicon in future versions.[32] The two efforts may prove complementary and, as a general policy, we intend to make our work compatible with UMLS knowledge sources wherever possible and practical.

General Conclusions

The problem we are addressing is a central challenge in medical informatics. While it has the superficial form of a "vocabulary" issue, it actually embraces all the concerns of medical concept modeling.

Individually, we have attempted to solve some facets of the problem of mapping language into computationally tractable representations. We have found no general solution; we have not been successful in using existing resources, such as standard vocabularies, for such purposes. Collaboratively, we have found a common philosophy and methodology and have begun to address the problem as though we were engaged in scientific discovery. We do not have all the answers, of course, to the complex problem of medical concept representation, but we believe we have identified the appropriate focus—on the concept model and the rules of concept formation and language interpretation.

The work we have described above is only just beginning. We believe that our methodology will insure that our results will have a scientific foundation and will be useful to people who have not participated with us in their development. We expect our task to be difficult; but, we remain convinced that the creation of a uniform representational basis for medical language is absolutely critical for our own work and that of others in the medical informatics community.

Members of the Canon Group who have collaborated in writing this paper include (alphabetically):

Douglas S. Bell, MD
Keith E. Campbell, MD
Christopher G. Chute, MD, DrPH
James J. Cimino, MD
David A. Evans, PhD
Carol Friedman, PhD
Robert A. Greenes, MD, PhD
William R. Hersh, MD
Stanley M. Huff, MD
Stephen B. Johnson, PhD
Robert C. McClure, MD
Mark A. Musen, MD, PhD
Edward Pattison-Gordon, MS
Alan Rector, MD, PhD
Roberto Rocha, MD

References ■

1. Blair DC, Maron ME. An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM. 1985;28:289–99.
2. Haimowitz IJ, Patil RS, Szolovits P. Representing medical knowledge in a terminological language is difficult. In: Greenes R, ed. Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care, November 6–9, 1988, Washington, DC. Silver Spring, MD: IEEE Computer Society Press, 1988:101–5.
3. United States National Center for Health Statistics. International Classification of Diseases, Ninth Revision, with Clinical Modifications. Washington, DC: The Center, 1980.
4. Côté RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L, eds. The Systematized Nomenclature of Medicine: SNOMED International. Northfield, IL. College of American Pathologists, 1993.
5. Gabrieli ER. A new electronic medical nomenclature. J Med Syst. 1989;13(6):355–73.
6. Saint Yves IF. The Read Clinical Classification. Health Bull (Edinb.) 1992;50(6):422–7.
7. Dunham G, Henson D, Pacak M. Three solutions to problems of categorizing medical terminology. Methods Inf Med. 1984; 23:87–95.
8. McMahon LF, Smits HL. Can Medicare prospective payment survive the ICD9-CM disease classification system. Ann Intern Med. 1986;104(9):562–6.
9. Cimino JJ, Hripcsak G, Johnson SB, Clayton PD. Designing an introspective, controlled medical vocabulary. In: Kingsland LW, ed. Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care, November 5–8, 1989, Washington, DC. Washington, DC: IEEE Computer Society Press, 1989:513–8.
10. Blois MS. Information and Medicine, The Nature of Medical Descriptions. Berkeley, CA: The University of California Press, 1984.
11. Evans DA, Gadd CS. Managing coherence and context in medical problem-solving discourse. In: Evans DA, Patel VL, eds. Cognitive Science in Medicine. Cambridge, MA: MIT Press, 1989.
12. Church A. Introduction to Mathematical Logic, I. Princeton, NJ: Princeton University Press, 1956.
13. Findler NV, ed. Associative Networks: Representation and Use of Knowledge by Computers. New York: Academic Press, 1979.
14. Brachman RJ, Fikes RE, Levesque HJ. KRYPTON: integrating terminology and assertion. Proc AAAI-83. 1983;3:31–5.

15. Brachman RJ, Levesque HJ. The tractability of subsumption in frame-based description languages. Proc AAAI-84. 1984; 4:34–7.
16. Levesque HJ, Brachman RJ. A fundamental tradeoff in knowledge representation and reasoning. In: Brachman RJ, Levesque HJ, eds. Readings in Knowledge Representation. Los Altos, CA: Morgan Kaufmann, 1985:42–70.
17. MacGregor R. The evolving technology of classification-based knowledge representation systems. In: Sowa J, ed. Principles of Semantic Networks: Explorations in the Representation of Knowledge. San Mateo, CA: Morgan Kaufmann, 1991: 385–400.
18. Sager N, Friedman C, Lyman MS, et al. Medical Language Processing. New York: Addison Wesley, 1987.
19. Evans DA. Final Report on the MedSORT-II Project: Developing and Managing Medical Thesauri. Technical Report No. CMU-LCL-87-3. Pittsburgh, PA: Laboratory for Computational Linguistics, Carnegie Mellon University, 1987.
20. Evans DA. Pragmatically-structured, lexical-semantic knowledge bases for unified medical language systems. In: Greenes R, ed. Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care, November 6–9, 1988, Washington, DC. Silver Spring, MD: IEEE Computer Society Press, 1988:169–73.
21. Rector AL, Nowlan WA, Kay S. Unifying medical information using an architecture based on descriptions. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care, November 4–7, 1990; Washington, DC. Los Alamitos, CA: IEEE Computer Society Press, 1990:190–4.
22. Rector AL, Nowlan WA, Kay S. Conceptual knowledge: the core of medical information systems. In: Lun KC, Degoulet P, Piemme TE, Rienhoff O, eds. MEDINFO 92. Amsterdam:North-Holland, 1992:1420–6.
23. Rector AL, Nowlan WA, Glowinski A. Goals for concept representation in the GALEN Project. In: Safran C, ed. Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care, November 1–3, 1993, Washington, DC. New York: McGraw-Hill, 1994:414–8.
24. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural language text processor for clinical radiology. JAMIA. 1994;1:161–74.
25. Friedman C, Cimino JJ, Johnson SB. A schema for representing medical language applied to clinical radiology. Submitted to JAMIA.
26. Huff SM, Warner HR. A comparison of Meta-1 and HELP terms: implications for clinical data. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care, November 4–7, 1990, Washington, DC. Los Alamitos, CA: IEEE Computer Society Press, 1990:166–69.
27. Masarie FE Jr, Miller RA, Bouhaddou O, Giuse NB, Warner, HR. An interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. Comput Biomed Res. 1991;24(4):379–400.
28. Evans DA, Miller RA. Final Task Report (Task 2)—Unified Medical Language System (UMLS) Project: Initial Phase in Developing Representations for Mapping Medical Knowledge: INTERNIST-I/QMR, HELP, and MeSH. Technical Report No. CMU-LCL-87-1. Pittsburgh, PA: Laboratory for Computational Linguistics, Carnegie Mellon University, 1987.
29. Evans DA, Ginther-Webster K, Hart M, Lefferts R, Monarch I. Automatic indexing using selective NLP and first-order thesauri. In: RIAO '91. Barcelona: Autonoma University of Barcelona, 1991:624–44.
30. Evans DA, Hersh W, Monarch I, Lefferts R, Handerson S. Automatic indexing of abstracts via natural-language processing using a simple thesaurus. Med Decis Making. 1991; 11(suppl):S108–15.
31. Lindberg DAB, Humphreys BL, McCray AT. The unified medical language system. In: van Bemmel JH, McCray AT, eds. 1993 Yearbook of Medical Informatics. The Netherlands: International Medical Informatics Association, 1993:41–51.
32. McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A, Srinivasan S. UMLS knowledge for biomedical language processing. Bull Med Libr Assoc. 1993;81(2):184–94.