The Practice of Informatics JAMIA

Review ■

Representing Thoughts, Words, and Things in the UMLS

KEITH E. CAMPBELL, MD, PHD, DIANE E. OLIVER, MD,
KENT A. SPACKMAN, MD, PHD, EDWARD H. SHORTLIFFE, MD, PHD

Abstract The authors describe a framework, based on the Ogden-Richards semiotic triangle, for understanding the relationship between the Unified Medical Language System (UMLS) and the source terminologies from which the UMLS derives its content. They pay particular attention to UMLS's Concept Unique Identifier (CUI) and the sense of "meaning" it represents as contrasted with the sense of "meaning" represented by the source terminologies. The CUI takes on emergent meaning through linkage to terms in different terminology systems. In some cases, a CUI's emergent meaning can differ significantly from the original sources' intended meanings of terms linked by that CUI. Identification of these different senses of meaning within the UMLS is consistent with historical themes of semantic interpretation of language. Examination of the UMLS within such a historical framework makes it possible to better understand the strengths and limitations of the UMLS approach for integrating disparate terminologic systems and to provide a model, or theoretic foundation, for evaluating the UMLS as a Possible World—that is, as a mathematical formalism that represents propositions about some perspective or interpretation of the physical world.

■ **JAMIA.** 1998;5:421–431.

> "When *I* use a word," Humpty Dumpty said in a rather scornful tone, "it means just what I choose it to mean—neither more nor less."
>
> "The question is," said Alice, "whether you *can* make words mean so many different things."
>
> "The question is," said Humpty Dumpty, "which is to be master—that is all."
>
> —Lewis Carroll

Affiliations of the authors: Stanford University, Stanford, California (KEC, DEO, EHS); Oregon Health Sciences University, Portland, Oregon and College of American Pathologists, Northfield, Illinois (KAS); Veterans Administration Health Care System, Palo Alto, California (DEO).

Correspondence and reprints: Keith E. Campbell, MD, PhD, Section on Medical Informatics, Stanford University School of Medicine, Stanford, CA 94305-5479.

This work was supported by training grant LM-07033 and resource grant LM-05305 from the National Library of Medicine and by a medical informatics postdoctoral fellowship from the Veterans Administration.

Received for publication: 2/3/98; accepted for publication: 4/17/98.

Many informatics developers have struggled to understand how best to leverage the Unified Medical Language System (UMLS) in their applications. One fundamental decision faced by these developers is whether to treat the UMLS as if it were a coding system unto itself—by using the UMLS Concept Unique Identifier (CUI) to represent concepts communicated in coded form and archived in data repositories—or whether to rely directly on one or more of the UMLS source terminologies for concept representation and use the UMLS as a system for providing appropriate interoperability between terminologies, as when trying to "facilitate the development of conceptual connections between users and relevant machine-readable information."[1]

We previously argued that evaluating the UMLS as if it were itself a coding system places it in a competitive position with the very sources from which it is derived and does not help us understand the unique value that the UMLS intends to provide.[2] Here we seek to characterize other consequences of treating the UMLS as if it were a coding system, by describing the limitations of communication inherent in our language and the implications that those limitations have on understanding the "meaning" of terms represented by a CUI.

Because the UMLS is founded on language, we present first a historical framework for understanding the relationships among the words we speak and write, the thoughts we are trying to express with our language, and the things to which our words and our thoughts refer. Next, we focus on the UMLS approach to concept representation and terminology specification, considering the UMLS in the context of this framework for thoughts, words, and meaning. By seeking a historical perspective, we intend to demonstrate the timeless nature of several principles of semantic interpretation and their applicability to the UMLS.

Thoughts, Words, and Things

How thoughts, words, and things* relate to one another has been a recurrent theme in scholarly works of philosophy and language from as early as Plato to the modern era. Plato dealt with the question of the proper naming of things in *Cratylus*, a dialog in which the participants argue over whether names are correct simply because they are used by convention (conventionalist view) or whether, in an ideal language, names would be most correct if they resemble or naturally describe the entities they name (naturalist view).[4] Plato himself seems to propose that neither of these views is completely accurate. In an optimal world, the purpose of names would be to ensure that a particular expression will make everybody think of one and only one thing. Plato, however, was doubtful that perfect names, which would reflect the character of the things they represent could ever be given, because things are continually changing. If things are continually changing, then there is no way to know what a thing is or what it is really like.

Aristotle went beyond the question of names and was interested in definitions. His notion of definition was not, as we usually think of it, simply the linguistic meaning of a word but was meant to explain clearly what a thing is by being a statement of the "essence" of the entity.[5] Aristotle believed that to say *what* something is, one must say *why* something is; therefore, his definitions were causal. An Aristotelian definition is given by specifying the genus and differentia of individuals and then using logical arguments to categorize those individuals on the basis of their definitions. By identifying the common definitional properties of similar individuals, the definition explains why they are members of the same kind. Representing terms relevant to health care data using an Aristotelian approach can provide a logical foundation for representing clinical data[6]; however, such an Aristotelian foundation is not sufficient. The language used to represent the terms of an Aristotelian system must be sufficiently precise to allow the terms to be reproducibly understood and applied. We cannot ignore, however, the unavoidable limitations of communicating meaning via language and the inherit ambiguities created by implicit exchange of different "senses" of meaning.

Understanding such ambiguities was a topic that concerned Gottlob Frege (1848–1925), a German philosopher and logician. Frege was not only one of the founders of mathematical logic and set theory but also a significant contributor to the philosophy of language.[7,8] Frege studied the meaning of proper names and concept words in his essay "On Sinn and Bedeutung," in which he distinguished between two types of meaning: thought content and referent. We look at Frege's ideas, particularly the two types of meaning he articulated, as a basis for understanding the semantic framework of the UMLS. Ogden and Richards popularized the importance of understanding the difficulties posed by these different senses of meaning and graphically illustrated the relationship between

*We borrow the phrase "thoughts, words, and things" from the title of the first chapter of Ogden and Richards' seminal book *The Meaning of Meaning*,[3] originally published in 1923.

language, thought content, and referent in a diagram commonly referred to as either the *semiotic triangle* or the *meaning triangle* (Figure 1).

The diagram shows that, although written or spoken symbols (words) cannot completely capture the essence of a reference (thought) or of a referent (thing), there is a correspondence among them. Either a word or an object can inspire a thought, and people may endeavor to express their thoughts with words or by identifying objects in the world. The relationship between a word and a thing is indirect, however. The link can be completed only when an interpreter (usually a person) processes the word, which invokes a corresponding thought, and then links that thought to a thing in the world (the "referent"). This diagram is seductive in its simplicity.

By implying a one-to-one relationship between each pair of members in the triangle, this simple diagram masks hidden complexity. Ogden and Richards alluded to this complexity by the dotted line between a symbol and a referent, indicating that the link between a symbol and a referent can only be made indirectly through an interpreter, but the notion that a symbol does—or could—refer to a single thought and that a thought does—or could—refer to a single referent is a fallacy. Many recognize that we live in a world where referential complexities lead to difficulties in communication. In Lewis Carroll's *Alice Through the Looking Glass* Humpty Dumpty insists that he can make words mean whatever he wants them to mean. In Plato's *Cratylus*, Socrates argues that it is not enough to try to understand what a thing is, based on its name, because the name-givers may have been living in ancient times, and the name reflects only what the name-givers thought was the nature of reality then; however, they may have been wrong. Thus, it has been historically recognized that multiple terms may refer to the same object or idea, a single term may refer ambiguously to more than one object or idea, and terms may be confusing because they are out of date. It is within this context that we seek solutions to improve our ability to communicate about biomedical concepts.

In professions where ambiguous communication can have deadly consequences, as in medicine, there is a strong desire to have one-to-one relationships between thoughts, words, and things among all the participants (interpreters) in the process. In a supplement to *The Meaning of Meaning* by Ogden and Richards, a physician named Crookshank recognized the clarity of communication that would result if medicine could develop an unambiguous relationship between thoughts, words, and things. He accordingly chastised the medical profession:

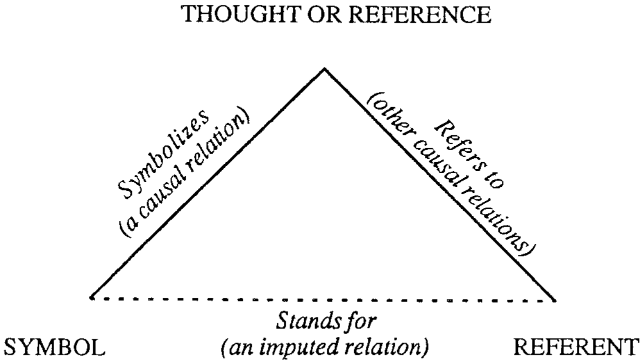
THOUGHT OR REFERENCE
Symbolizes (a causal relation)
Refers to (other causal relations)
SYMBOL
Stands for (an imputed relation)
REFERENT

Figure 1 The Ogden and Richards semiotic triangle, from the original 1923 illustration.[3]

> Medicine . . . [has] forfeited pretension to be deemed a Science, because her Professors and Doctors decline to define fundamentals or to state first principles, and refuse to consider, in express terms, the relations between Things, Thoughts and Words involved in their communication to others.[9]

Crookshank's goal of improving communication is laudable and has proved effective in the "exact sciences" and in the legal world (for example, in the preparation of formal contracts as discussed by Eco[10]). Such precision is an unobtainable goal for medicine, however, if for no other reason than the imperfectability of human beings and the huge regional variations in disease and its manifestations. Even more pertinent, of course, is the magnitude of medical knowledge that must be mastered to approach the understanding necessary to delineate completely all the words used in the profession, all the possible thoughts that might be invoked by those words, and all the possible things to which those thoughts might refer.

Another factor that complicates our efforts to improve communication is the vertical nature of medicine, as described by Blois.[11] He points out that medical knowledge (and hence communication regarding medicine) is vertically organized, in that mastery of the discipline requires a corpus of knowledge that ranges from the foundational and relatively exact sciences of physics and chemistry (which have precise symbolic mechanisms of communication such as the periodic table of the elements) to psychology and sociology (where the elements of discourse are often intangible and difficult to communicate). Hence, our ability to be precise in medical communication is challenged as one goes "up" this vertical scale from basic sciences to the social and psychological milieu in which a patient functions and is assessed.

Although we argue that the complete delineation of the thoughts, words, and things relevant to medicine, and their subsequent encoding, are laudable but elusive goals, we readily concede that the pursuit of such a codification, such as the UMLS is undertaking, will improve the clarity with which we communicate. To optimize the utility of such a codification, however, it is critical that we understand its limitations. Unrecognized ambiguity created by interchanging different "senses" of meaning is one such limitation that we discuss in the following sections.

Extensional and Intensional Meaning

In a classic example by Frege, the names "morning star" and "evening star" are expressions that refer indirectly to the same physical object, the planet Venus. Although today we know this is the case, there was a time when people were not aware of the correspondence between the "physical objects" implied by the two terms. Our interpretation of the world is shaped by our experiences, and they in turn determine how we communicate with one another. Our ability to understand one another depends on having sufficient shared experiences that we can invoke common thoughts when other people confront us with appropriate words and things. Although in one sense we can say that "morning star," "evening star," and "Venus" are equivalent (that is, they have the same meaning in the sense that they all refer to the same planet), one can also say that "morning star," "evening star," and "Venus" are not equivalent (in the sense that more information is connoted by these names than simply the physical objects to which they refer, such as when the entity can be observed and, perhaps, the experiences of the observer). In such statements, Frege recognized a puzzle: How can we say the same thing about the same objects but mean different things? Consider the following two expressions:

> The morning star is low in the sky.
> The evening star is low in the sky.

These expressions refer to the same object (Venus), yet the sense of meaning conveyed by the two expressions is very different (consider the implicit time of day conveyed by the expressions). Frege explained this puzzle by recognizing that the "meaning" of expressions can be divided into two components: On the one hand there are the physical objects to which the expression refers (the expression's *extensional* component) and on the other there are the characteristic features of the physical object used to identify it (the expression's *intensional* component).† Understanding the interrelationship between intensional and extensional meaning is essential to understanding the "senses" of meaning represented within the UMLS. Only with this understanding can we know when symbols (such as morning star and evening star) can be substituted for one another without loss of truth.

Figure 2 uses the Ogden-Richards triangle to illustrate the difference between the intension and extension of an expression. The intension is the *connotation*, and the extension is the explicit *denotation*. An individual's experiences will determine how he or she will interpret expressions. Imagine someone who has seen the sun rise with the morning star, another who has never seen the sun rise but who knows that the morning star refers to the second planet, and a third individual, an astronomer perhaps, who has personally observed the morning star and the evening star and also knows that both terms refer to the second planet. Figure 2 illustrates the different thoughts that the three expressions might evoke in each of these persons' minds. In the first person, mention of the morning star might evoke a vivid memory of how the star looked when last gazed upon. In the second, it might evoke the orbits of the planets in the solar system and the position of the planet Venus with respect to the sun and the other planets. In the last, mention of any of the three (morning star, evening star, or Venus) might evoke any of these thoughts, depending on the context in which mention is made. Yet we realize that, in the corporeal world, *all* these thoughts are linked to the planet Venus. As Figure 2 illustrates, the thoughts invoked by words are dependent on our individual backgrounds, and opportunities for referential complexity abound. The division of meaning into intensional and extensional components, and the coupling of our interpretation based on our experiences (the *intensional* meaning) is a fundamental limitation of the systematization of communication one derives from the Ogden-Richards triangle.

The UMLS as a Possible World

The Ogden-Richards triangle has elucidated the relationship between thoughts, words, and things in the corporeal world by providing a conceptual framework that we can use to better understand our use of the language. The UMLS is an example of an artifact

†Jaroslav Peregrin credits Carnap with replacing Frege's distinction between *Sinn* and *Bedeutung* (Frege's original German words, which have been translated as "sense" and "denotation") with the distinctions between intension and extension.

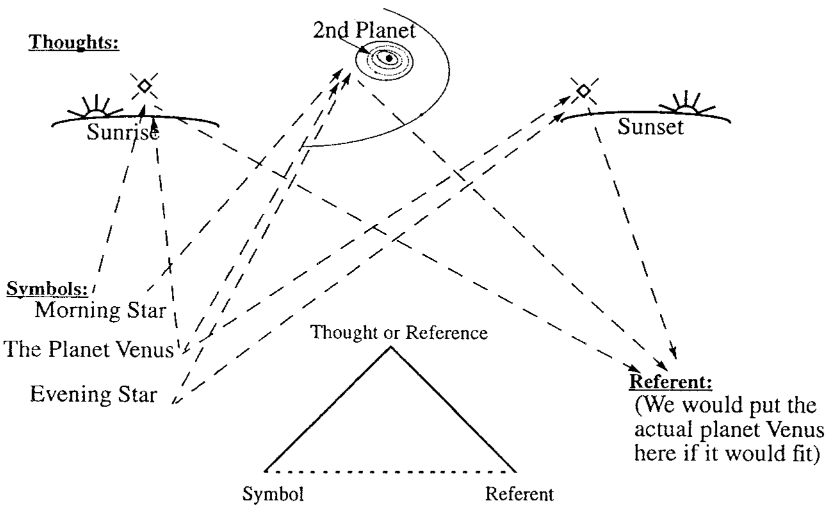

Figure 2 The Ogden and Richards semiotic triangle applied to the notions of three expressions that all refer to the same physical object but generate different intensional thoughts. Notice that there is not a one-to-one correspondence between the thoughts, words, and things represented by the three corners of the meaning triangle.

that embodies these same relationships in what has been called a Possible World, an artificial system in which relationships can be formally codified and the truth values of these relationships can be evaluated.‡ The notion of Possible Worlds originated with the 17th-century German philosopher Gottfried Leibniz (1646–1716), who used the notion for theologic purposes. Leibniz said that God could create only logically possible worlds, but that being omniscient and beneficent, he would actualize the best of all possible worlds.[14] Modern logicians have abandoned the theologic implications in employing the idea of Possible Worlds. One approach taken by formal semanticists, dating back at least to the work of the logician Jørgen Jørgensen in the 1930s,[15] is to say that a Possible World is simply a set of propositions. This approach has subsequently been much discussed in the philosophy and artificial intelligence literatures (see, for example, Herbert Simon's classic book *The Sciences of the Artificial*[16]). In describing the UMLS as a Possible World, we state that it is possible that there exists a perspective on the world where the correspondence among words, thoughts, and things are exactly as represented by propositions within the UMLS.§

We also claim that the UMLS contains all the characteristics of the Ogden-Richards triangle. This should not be surprising, since the approach taken in developing the UMLS was founded on the belief that the essential properties of biomedicine necessary to construct the UMLS would reveal themselves via the properties of language used to describe the discipline,[17] and the Ogden-Richards triangle was developed to explain the relationship of language to thoughts and to the world. Thus, the development of the UMLS was empirically driven. Rather than attempting the creation of a system de novo, the developers collected the language that others had codified into terminologic systems, provided a framework where the intension (connotation) of terms of those systems could be preserved, and unified those systems by providing a representation of extensional meaning|| by collecting abstract concepts into sets that can be interpreted to represent their extension.¶

These extensional sets are codified by the *Concept Unique Identifier* (CUI) in the UMLS. We argue that the "meaning" of this identifier is only understandable extensionally, by examining the characteristics shared

‡We do not seek in this article to prove formally the analogy between the UMLS and Possible Worlds as used in modal propositional calculi. However, we believe that the UMLS can readily fit into formal Possible World frameworks such as described by Stein,[13] and that such a framework may provide useful insights and perhaps functionality.

§The UMLS contains propositions that state which terms are equivalent to one another. Each such proposition thereby specifies the existence of a concept and links that concept to a set of terms from the source terminologies that are equivalent given a particular perspective represented by a Possible World.

||The meaning of a term in the extensional sense is given just by listing, or somehow indicating, what things are referred to by the term.

¶For philosophers prior to Frege the extension of a concept was reserved for only physical objects that share the essential characteristics of a concept. Frege introduced the notion that the extension of a concept can also refer to abstract concepts that share the essential characteristics of a concept. For example, Frege used such extensional sets to represent a definition of the "direction of line *a*" as the extension of the concept "parallel to line *a*."[8]

Table 1 ■

UMLS Phrases Corresponding to the Concept C0004057 named "Aspirin"

Aspirin	Ecotrin	Entericin
aspirin	Endosprin	St. Joseph
Acetylsalicylic Acid	Magnecyl	Measurin
Acetylsalicylic acid	Micristin	ACIDE ACETYLSALICYLIQUE
acetylsalicylic acid	Polopiryna	ACETYLSALICYLIQUE, ACIDE
Acid, Acetylsalicylic	Zorprin	ASPIRINE
Acetysal	2-(Acetyloxy)benzoic Acid	ASPIRIN
Acylpyrin	Benzoic acid, 2-(acetyloxy)-	ASPIRINA
Colfarit	Aspergum	ACIDO ACETILSALICILICO
Easprin	Empirin	

NOTE: Notice the differences in word order and capitalization among the individual phrases.

by all abstract concepts linked by a CUI. We will illustrate the correspondence of the UMLS and the Ogden-Richards triangle by using the notion of "aspirin" as a prototypical example in the following sections.

Words

Words represent written and spoken communication by which we convey meaning to one another. In an "artificial" world, such as the UMLS, words provide a link between the realm in which we live and the symbolic world in which computer programs operate. As such, the notion of words that are part of our language expands to include the notion of symbols that represent a source terminology's "representation," such as a Systematized Nomenclature of Medicine (SNOMED) term code.[18]

Table 1 presents phrases in the UMLS that correspond to the concept C0004057 named Aspirin. Note that the entries include terms in several languages, since the UMLS contains entries in English, French, German, Spanish, and Portuguese. Although the initial set of phrases came from the source terminologies, UMLS provides added value by cataloging preferred forms and categorizing lexical variants by word order and by case differences.

Thoughts

When developers of source terminologies developed their systems, they had very specific thoughts about what the individual terms "meant" (in the intensional sense) with respect to the terminology they were developing and the human beings who would interact with those systems. Although we cannot directly know what was in the minds of the developers of the source terminologies, the UMLS developers have used clues embodied within the sources to try to infer what those thoughts were and to try to codify those thoughts within the UMLS. These clues take several forms: the term used by a source to describe the thought; the synonyms used by a source to describe other statements that its developers considered equivalent to the thought; and any formal or informal relationships used by the developers to relate terms within the terminologic system to one another. Some of the informal relationships had to be inferred from processing the typesetting tapes for a particular source, using constructs such as how many tabs appeared before the word, whether the word was in bold or italics, and what page of the printed book the word occurred on.

Figures 3 and 4 show how two different UMLS sources, SNOMED International and the Computer Retrieval of Information on Scientific Projects (CRISP) Thesaurus,[19] represent the terms and relationships that the UMLS have determined are equivalent to the concept "Aspirin" (by virtue of sharing the CUI C0004057). In the case of SNOMED, three terms that were assigned the same CUI by the UMLS were felt to be in some respect different by SNOMED, since they have unique SNOMED term codes. Other potential differences in the interpretation of the meaning intended by a source are evident by noting how the hierarchies used to classify a term in each source differ in their granularity and in some cases in their organizing principles.

It is obvious that the intension associated with a term in a source terminology is represented at least in part by its location in a hierarchy and by decisions made regarding synonyms and nonsynonyms. Aspirin in the CRISP Thesaurus is a chemical; it is also a centrally acting drug that has antirheumatic, anti-inflammatory, analgesic, and antipyretic properties. Similarly, the UMLS equivalent of aspirin in SNOMED, acetylsalicylic acid, is a chemical. It is also a drug with several of the same properties that it has in the CRISP Thesaurus: It is a centrally acting agent, an analgesic, and an antipyretic. On the other hand, in SNOMED,

CRISP Thesaurus

Identifier	Term
2270-2387	aspirin

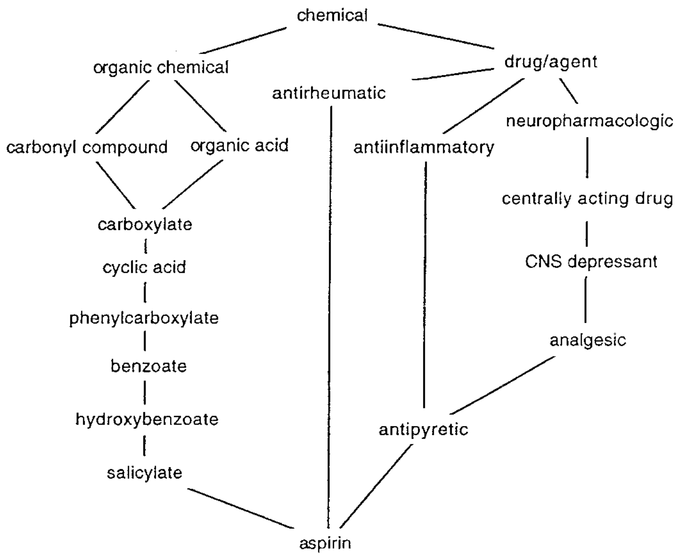

Figure 3 A term from the CRISP Thesaurus that is part of the set of terms designated by the UMLS CUI C0004057 named "Aspirin," and relationships from the CRISP Thesaurus that classify that term. See Table 1 for a complete list of phrases associated with CUI C0004057.

SNOMED International

Identifier	Term
C-C17E6	Easprin
C-C29EA	Zorprin
C-60320	Acetylsalicylic acid

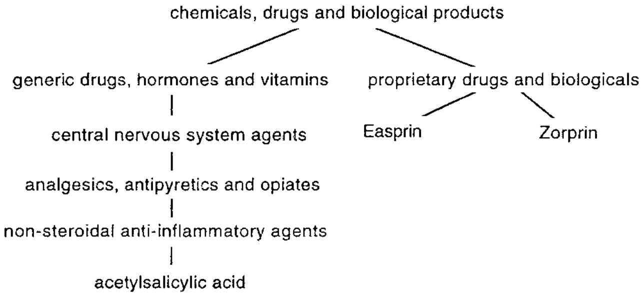
chemicals, drugs and biological products
generic drugs, hormones and vitamins
proprietary drugs and biologicals
central nervous system agents
Easprin
Zorprin
analgesics, antipyretics and opiates
non-steroidal anti-inflammatory agents
acetylsalicylic acid

Figure 4 Terms from SNOMED that are part of the set of terms designated by the UMLS CUI C0004057 named "Aspirin," and relationships from SNOMED that classify those terms. See Table 1 for a complete list of phrases associated with CUI C0004057.

acetylsalicylic acid is not synonymous with two other UMLS equivalents of aspirin, Easprin and Zorprin, because the first is a generic drug and the other two are proprietary drugs. Thus, in SNOMED, the intension of aspirin is clearly not the same as the intension of Easprin, yet aspirin and Easprin are linked to the same CUI. It may even be argued that there are subtle differences in the intension of aspirin in CRISP and SNOMED, yet these differences are obscured or lost when one moves from the source terminology to the CUI.

It is clear that the intensional meanings, or connotations, of terms in the different sources are distinct. Some source terminologies, such as SNOMED, can have very precise semantics, differentiating between the observation of a particular pathologic change (such as a fracture) and the diagnostic statement that a fracture may exist in a particular patient affecting a particular bone. SNOMED maintains these distinctions by maintaining independent hierarchies (such as topography, morphology, procedures, and diagnoses) that embody different intensions. Other terminologies, such as the National Library of Medicine's Medical Subject Headings (MeSH),[20] have no need for such detailed specificity, and thus the intensional meanings of similar phrases in MeSH cannot be interpreted in the same way that they might be in SNOMED. For example, "gastrointestinal transit" in MeSH is used to denote both

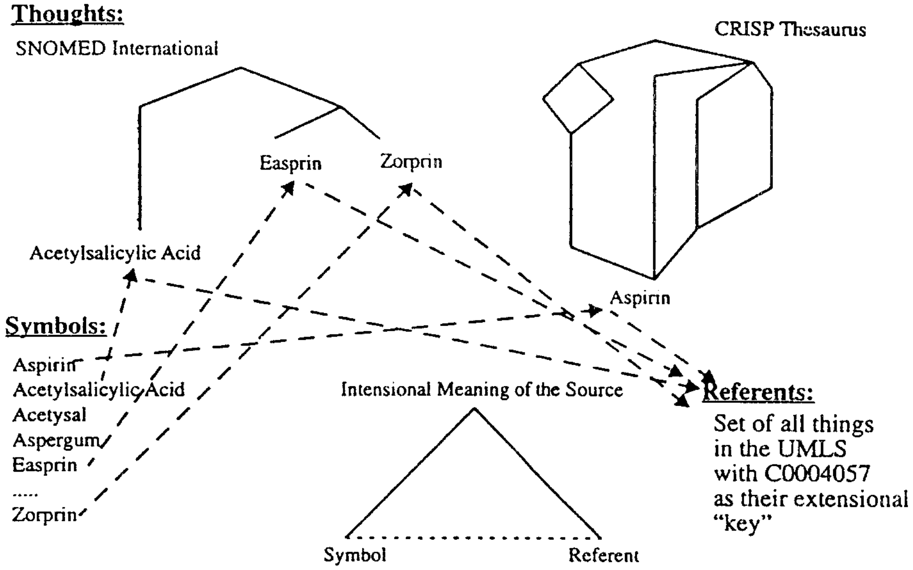

Figure 5 Integration of symbols representing aspirin, intensional meanings of aspirin, and extensional referents in two of the UMLS sources. Compare with Figure 2.

the physiologic function and the diagnostic measure.[21] SNOMED would use separate codes to differentiate these two notions, by putting a "gastrointestinal transit" term that represented physiologic function into the function axis and putting a "gastrointestinal transit" term into the procedure axis that represented the act of performing the diagnostic measure.

The point here is not that the intensional meanings of phrases in SNOMED are somehow more correct or desirable than the intensional meanings of phrases in MeSH. MeSH and SNOMED are different because they were intended for different purposes. Thus, for the UMLS to meet its goal of integrating sources such as SNOMED and MeSH into a useful framework, it must represent the different intensions of the sources while also providing appropriate integration.

Developers of the UMLS recognized this requirement to represent the notion of meaning relative to the scope, granularity, context (hierarchy), synonyms, and annotations of the source terminologies,[21] and we believe that this aspect of their design has been successful. For each source, the UMLS explicitly represents the hierarchic context in which the terms are encountered and links those terms to other "extensionally equivalent" terms (in this case, all terms linked to the referent "Aspirin") from all the UMLS sources. We discuss these extensional representations in the next section.

Things

To complete the Ogden-Richards triangle, the UMLS must have a representation of "things." A museum may represent things by collecting example "prototypical" artifacts from around the world and making them available for inspection and experimentation. For the UMLS, the task is somewhat more complicated, since many of the "things" it must represent have no physical manifestation and are comprehensible only in the abstract. Thus, the UMLS uses an abstract notion of "concepts" to represent classes of "things" that can in some sense be considered equivalent, and it provides a CUI as a means of codifying the extension of these classes. It can be argued that the complete notion of what these CUIs represent is understandable only extensionally, as the characteristics shared by all the intensional representations linked together via a common CUI, including both relationships or annotations derived directly from UMLS sources themselves as well as other relationships or annotations that are provided during UMLS construction.

Figure 5 represents this intensional (terms and relationships provided by a particular source for a particular purpose) and extensional codification for the UMLS concept "C0004057," which corresponds to some interpretation of the meaning of "Aspirin." Close examination of Figure 5 may raise questions. One is whether the extensional sets (i.e., terms, semantic types, definitions, and relationships that share the same CUI) of the UMLS are appropriate for the intended purpose of the UMLS. Table 2 presents such a set for the concept C0004057.

In Figure 5 (as in the UMLS), aspirin, acetylsalicylic acid, Easprin, and Zorprin exist as intensionally distinct concepts in at least one source but are all linked to the same extension (i.e., the set of terms that share the same CUI). Sources not presented in Figure 5 have additional terms, such as Aspergum and Ecotrin, that are included in the extensional set represented by C0004057 (the complete list appears in Table 1). We readily concede that this equivalence may be true in some Possible World, but the question is how appropriate this particular extensional representation is for the purposes of the UMLS (which seeks "to facilitate the development of conceptual connections between users and relevant machine-readable information"). Many clinicians would not regard different formulations of aspirin, Ecotrin (an enteric-coated aspirin), and Aspergum (a chewing-gum preparation of aspirin) as interchangeable concepts in the prescriptions they write. Although aspirin may be an abstract concept, Ecotrin and Aspergum have specific formulations (extensions) in our corporeal world, and use of those particular formulations is subject to different indications, mechanisms of therapy, and risks to the patient. Clearly then, in at least a pharmacy order-entry system, any extensional relationship that was used to determine allowable substitution of pharmacologic formulations would need to have different relationships (representing a different Possible World), than the one currently embodied within the UMLS. However, for a system primarily concerned with the active ingredients of a drug, such as an allergy or drug interaction application, the Possible World embodied in the UMLS may be optimal.

Table 2 ■

Extension of the Concept C0004057 Named "Aspirin" in the UMLS

Semantic types:

Organic Chemical
Pharmacologic Substance
Biologically Active Substance

Definitions:

- **Aspirin**: Acetylsalicylic acid, a drug having anti-inflammatory, analgesic, and antipyretic effects; it is the prototype of the nonsteroidal anti-inflammatory agents whose mechanism of action is inhibition of prostaglandin synthesis; used for relief of pain and fever, for treatment of rheumatoid arthritis and osteoarthritis, and for antiplatelet therapy to reduce the risk of recurrent transient ischemic attacks or of cerebrovascular accident. (From Dorland's Dictionary, 27th ed.)
- **Ecotrin**: Trademark for a preparation of aspirin. Dorland's Illustrated Medical Dictionary 27th edition. (From Dorland's Dictionary, 27th ed.)
- **Aspirin**: The prototypical analgesic used in the treatment of mild to moderate pain. It also has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase, which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. (From Martindale, *The Extra Pharmacopœia*, 30th ed.)

Phrases: See Table 1

NOTE: The UMLS relationships that are linked to concept C0004057 have been omitted from the table to simplify the presentation, but they are crucial to an understanding of the concept. Some of the relevant relationships are shown in Figures 3 and 4. Notice that the three definitions linked to C0004057 differ, even to the point that Dorland's Dictionary, from which two definitions were acquired, provides two separate entries.

This observation does not mean that we believe that the focus of representation in the UMLS is incorrect, nor do we mean that this focus is incapable of evolving to meet the needs of the consumers of the UMLS—quite the contrary. Because the representational framework of the UMLS is consistent with the historical themes of interpretation and meaning, and because it provides a framework for linking thoughts, words, and things relevant to the medical domain into a codified system that represents a Possible World, we can begin to have a dialogue about our intensional and extensional representational needs. We can thereby determine whether the current Possible World is an appropriate embodiment of our needs or whether the content of the UMLS needs to evolve in a particular way to better meet our needs. By providing a formal framework where we can begin to ask these questions of specific concepts, the UMLS enables a dialogue about the meaning of "aspirin" that we could not have had without the creation of a Possible World as a starting point. Winograd refers to such formal frameworks as systematic domains,[22] a structured formal representation that provides precise and unambiguous description of the tasks (the process of assigning "meaning" in the case of the UMLS) and forms the basis for tools that aid in communication and the cooperative accumulation of knowledge.

We can begin cooperative accumulation of knowledge through meaningful evaluation of the UMLS, but the evaluation must be appropriate for its framework, content, and purpose. Through such an evaluate-and-revise cycle, we can approach a codification of medical language that will improve the clarity with which we communicate, even if it is impossible to achieve a codification that is entirely unambiguous.

Appropriateness of a Possible World

Terms that share the same CUI are equivalent in a particular Possible World—that is, in the Possible World represented by the UMLS. Any evaluation of the UMLS should include an evaluation of the appropriateness of this Possible World. However, evaluators can perform such a study well only if they have an understanding of the Possible World the UMLS intends to support.

In our previous discussion of how the UMLS represents "Aspirin," we noted that different formulations of aspirin (aspirin, Aspergum, and Ecotrin) has been represented as distinct intensional meanings in one of the sources (see Figure 5), yet were linked with the same extensional meaning during UMLS construction. In that discussion we noted that most clinicians would probably not consider these three concepts interchangeable in the prescriptions they write. However, we also assert that from some possible perspectives, such as when we are concerned primarily with medication allergies, having these concepts all linked to the same extension makes perfect sense.

As another example, the UMLS CUI "C0002871" groups the SNOMED term *Anemia, NOS*# and the ICD-9-CM term *Anemia, unspecified* together. On the surface, this mapping appears to be quite reasonable. However, there is a problem that occurs when we consider iron deficiency anemia, which is clearly a kind of anemia. In ICD-9-CM, due to an explicit exclusion, iron deficiency anemia is excluded from the category *Anemia, unspecified*. Therefore, SNOMED's *Anemia, NOS* has a different intension than ICD-9-CM's *Anemia, unspecified* in that it includes iron deficiency anemia but the ICD term does not. Furthermore, if one mapped through the CUI from SNOMED to ICD-9-CM (or vice versa), a semantic inaccuracy would be introduced that could be undesirable in a medical record but might be beneficial in other settings (for example, by improving query recall when trying to link concepts from medical records to the medical literature).

What, then, is the right perspective, or the right combination of perspectives (to the extent that multiple perspectives can coexist), on extensional meaning for the UMLS? We need to characterize more explicitly the precise purposes for which the current Possible World of the UMLS is most appropriate. Through such characterization, we can not only delineate the proper perspectives for the UMLS but can also begin to ask questions about the proper granularity of concept representation the UMLS and its source terminologies should embody as well.

Discussion

In this paper, we have presented a framework for interpreting the semantics of the UMLS, paying particular attention to the different senses of meaning represented by the CUI and by the UMLS sources. We have tried to justify our interpretation of the UMLS by analogy, through use of Ogden and Richard's meaning triangle, and by demonstrating consistency with historical themes of semantic interpretation. Others have discussed Ogden and Richard's meaning triangle in the medical informatics context,[23–28] but none have used Ogden and Richard's framework as a direct aid for understanding the semantics of the UMLS. In addition, we promote the notion that the CUI is understandable only *extensionally*, by examining the terms from the UMLS sources that are linked via a CUI. Through this linkage, we argue that the CUI takes on *extensional* meaning—meaning that is different from the *intensional* meaning represented by the UMLS sources. This extensional interpretation of the UMLS—where we use a Possible World framework as the source for a CUI's extensional set, rather than the more typical corporeal world—is unique. An important implication of assigning the CUI extensional meaning is the recognition that a CUI's meaning changes any time a new term is included in—or an existing term is removed from—the CUI's extensional set. This change in meaning is not limited to identification of historical mistakes in the assignment of source terms to CUIs, but rather is part of the natural evolution of the semantics of the UMLS as new sources are integrated.

Some readers may suggest the merits of assigning intensional meaning to the CUI itself, in addition to its extensional meaning. In such a scenario, the CUI would represent both a class of concepts (all the terms from sources linked by a common CUI) and the extension of a higher order concept (a concept that must somehow be inferred by examining the class). Others may argue that such a scenario seems both incongruous and circular. As authors, we have not resolved this debate, even among ourselves.

Although we may debate the intensional and extensional characteristics that we individually assign to the CUI—based on the characteristics that the UMLS embodies today—the UMLS is an evolving artifact that may change in ways that will force us to reconsider our interpretation. Debating these intensional and extensional characteristics serves to highlight the ambiguity of what the CUI is, and what it is intended to be. Within such debate we must also consider the implications of various interpretations and use the understanding thus attained to inform the evolution of the UMLS in ways we collectively would consider optimal for the range of purposes for which the UMLS is likely to be used.

One implication of such debate is the need to consider the effects that various interpretations and uses of the UMLS CUI will have on the relationships among the

#NOS is an abbreviation for "not otherwise specified."

National Library of Medicine, the UMLS source providers, and vendors of terminology-enabled applications. There are important accountability, resource, contractual, copyright, and intellectual issues that must be considered. We hope that the UMLS framework will be understood in a way that fosters collaboration between the National Library of Medicine and the providers of the terminology systems it incorporates. With such an understanding, vendors of terminology-enabled applications are more likely to recognize synergistic value provided by both the UMLS and the source terminologies it integrates.

What direction will the relationships, accountability, and resource commitments surrounding our terminologic needs ultimately take? Although we cannot be certain, we certainly encourage an active debate. Through such debate our understanding of the problems will improve and our ability to solve those problems will be enhanced.

The authors thank Mark Tuttle, Stephanie Lipow, Alexa McCray, and John Thomas for thoughtful discussions and advice regarding this manuscript.

References ■

1. Humphreys BL, Lindberg DAB. The UMLS Project: making the conceptual connection between users and the information they need. Bull Med Libr Assoc. 1993;81(92):170–7.
2. Campbell KE, Oliver DE, Shortliffe EH. The Unified Medical Language System: toward a collaborative approach for solving terminologic problems. J Am Med Inform Assoc. 1998;5(1):12–6.
3. Ogden CK, Richards IA. The Meaning of Meaning. 8th ed. Orlando, Fla.: Harcourt Brace Jovanovich, 1989. First edition published 1923.
4. White NP. Plato on Knowledge and Reality. Indianapolis, Ind.: Hackett Publishing Co, 1976.
5. Witt C. Substance and Essence in Aristotle: An Interpretation of Metaphysics VII–IX. Ithaca, NY: Cornell University Press, 1989.
6. Campbell KE, Das AK, Musen MA. A logical foundation for representation of clinical data. J Am Med Inform Assoc. 1994;1(3):218–32.
7. Beany M (ed). The Frege Reader. Malden, Mass.: Blackwell, 1997.
8. Kenny A. Frege: an Introduction to the Founder of Modern Analytic Philosophy. London, England: Penguin Group, 1995.
9. Crookshank FG. The importance of a theory of signs and a critique of language in the study of medicine. In: Ogden CK, Richards IA. The Meaning of Meaning. 8th Ed. Orlando, Fla.: Harcourt Brace Jovanovich, 1989:337–55.
10. Eco U. Introduction: the meaning of "The Meaning of Meaning." In: Ogden CK, Richards IA. The Meaning of Meaning. 8th ed. Orlando, Fla.: Harcourt Brace Jovanovich, 1989:v–xi.
11. Blois MS. Medicine and the nature of vertical reasoning. N Engl J Med. 1988;318(13):847–51.
12. Carnap R. The Logical Foundations of Probability. Chicago, Ill.: University of Chicago Press, 1950.
13. Stein LA. Extensions as Possible Worlds. In: Sowa JF (ed). Principles of Semantic Networks: Explorations in the Representation of Knowledge. San Mateo, Calif.: Morgan Kaufmann Publishers, 1991:267–81.
14. Lyons J. Linguistic Semantics: An Introduction. Cambridge, England: Cambridge University Press, 1995.
15. Jørgensen J. Imperatives and logic. Erkenntnis. 1937–38;7: 288–96.
16. Simon HA. The Sciences of the Artificial. 2nd ed. Cambridge, Mass.: MIT Press, 1981.
17. Tuttle MS, Blois MS, Erlbaum MS, Nelson SJ, Sheretz DD. Toward a biomedical thesaurus: building the foundation of the UMLS. Proc 12th Annu Symp Comput Appl Med Care. 1988:191–5.
18. Côté RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L (eds). The Systematized Nomenclature of Medicine: SNOMED International. Northfield, Ill.: College of American Pathologists, 1993.
19. Division of Research Documentation. Computer Retrieval of Information on Scientific Projects (CRISP) Thesaurus. Bethesda, Md.: National Institutes of Health, 1997.
20. National Library of Medicine. Medical Subject Headings. Bethesda, Md.: NLM, 1992. Publication NTIS/NLM-MED-92-01.
21. McCray AT, Hole WT. The scope and structure of the first version of the UMLS semantic network. Proc 14th Annu Symp Comput Appl Med Care. 1990:126–30.
22. Winograd T, Flores F. Understanding Computers and Cognition: A New Foundation for Design. Reading Mass.: Addison-Wesley, 1986.
23. Baud R, Lovis C, Rassinoux A-M, Scherrer J-R. Alternate ways for knowledge collection, indexing and robust language retrieval. In: Chute CG (ed.) Proceedings of the IMIA Conference on Natural Language and Medical Concept Representation; Jacksonville, Florida. 1997:81–93. Also: Methods Inf Med, in press.
24. Scherrer J-R. Concepts, knowledge and language in healthcare information systems: followup 30 months later. In: Chute CG (ed). Proceedings of the IMIA Conference on Natural Language and Medical Concept Representation; Jacksonville, Florida. 1997:5–8. Also: Methods Inf Med, in press.
25. Rector AL. Thesauri and formal classifications: terminologies for people and machines. In: Chute CG (ed). Proceedings of the IMIA Conference on Natural Language and Medical Concept Representation; Jacksonville, Florida. 1997:183–95.
26. Ingernerf J, Giere W. Concept oriented standardization and statistics oriented classification: continuing the classification versus nomenclature controversy. In: Chute CG (ed). Proceedings of the IMIA Conference on Natural Language and Medical Concept Representation; Jacksonville, Florida. 1997:147–67. Also: Methods Inf Med, in press.
27. Ingenerf J. Taxonomic vocabularies in medicine: the intention of usage determines different established structures. Medinfo. 1995:136–9.
28. Ceusters W, Beukens F, De Moor G, Waagmeester A. The distinction between linguistic and conceptual semantics in medical terminology and its implications for NLP-based knowledge acquisition. In: Chute CG (ed). Proceedings of the IMIA Conference on Natural Language and Medical Concept Representation; Jacksonville, Florida. 1997:71–9. Also: Methods Inf Med, in press.