The Epistemological Foundations of Knowledge Representations

Elaine Svenonius

Abstract

This paper looks at the epistemological foundations of knowledge representations embodied in retrieval languages. It considers questions such as the validity of knowledge representations and their effectiveness for the purposes of retrieval and automation. The knowledge representations it considers are derived from three theories of meaning that have dominated twentieth-century philosophy.

The discipline of philosophy impacts other knowledge disciplines, particularly in the theoretical constructs they employ. The purpose of this paper is to explore how epistemology, that branch of philosophy concerned with how and what we know, has contributed to the design of knowledge representations embodied in retrieval languages designed for organizing information. Different retrieval languages make different presuppositions about what is meant by knowledge. These differences give rise to questions such as

- How valid are the knowledge representations embodied in different retrieval languages, i.e., how well do they do what they purport to do, i.e., to represent knowledge?
- How effective are they in facilitating the achievement of the objectives of a retrieval language: collocation, discrimination, and navigation?
- How amenable are they to automation and semantic interoperability?

In the course of the twentieth century, the problem of what and how we know has been dealt with through language analysis and theories of meaning. Three theories of meaning are especially relevant to the discussion of

Elaine Svenonius, UCLA Graduate School of Education and Information Studies, Moore Hall Box 951521, 405 Hilgard Ave., Los Angeles, CA 90095-1521

LIBRARY TRENDS, Vol. 52, No. 3, Winter 2004, pp. 571–587

knowledge representations: Operationalism, the Referential or Picture theory of meaning, and the Contextual or Instrumental theory of meaning.

Operationalism

Operationalism is a theory of meaning emanating from the philosophy of logical positivism. Logical positivism, an extreme form of empiricism, dominated philosophy of science in the first decades of the twentieth century. Empiricism holds that all knowledge is derivable from experience, i.e., from sense perceptions. For instance: our knowledge of time as used as a variable in a mathematical equation, e.g., $v = d/t$, is ultimately derivable from propositions recording our sensory experience of time. The experience upon which knowledge is based must be objective. This condition is expressed by the Principle of Verifiability, which states that in order to be meaningful, a proposition must be capable of verification. The totality of knowledge consists of all meaningful propositions. Examples of nonmeaningful propositions are those of an ethical, religious, or "esthetic kind," e.g., "truth is beauty" is not meaningful because it cannot be verified, therefore, it is excluded from the corpus of knowledge.

For a proposition to be verified, the concepts within it need to be defined operationally, i.e., they need to be defined constructively. In practice, defining a concept operationally often means defining it as a variable. Defining concepts as variables enables a discipline to advance. The most celebrated example of this phenomenon is Einstein's use of operational definitions in his analysis of simultaneity (Bridgman, 1938, p. 7). A graphic example of the practicality of operational definitions is that of Eddington's elephant sliding down a hill of wet grass (Eddington, 1929, pp. 251 ff). Eddington asks us to consider the mass of this sliding elephant. Conceivably it could be regarded as a property of the elephant ("a condition which we vaguely describe as 'ponderosity'") (p. 251); on the other hand, it could be regarded as a pointer reading on a scale, i.e., two tons. It may be intuitive to think of mass as a property, but Eddington observes: "we shall not get much further that way; the nature of the external world is inscrutable, and we shall only plunge into a quagmire of indescribables" (p. 251). He goes on to argue that it is more productive to regard mass as a pointer reading, i.e., as a value of a variable. Not only does this give a method for testing the proposition "the elephant weighs two tons"; it enables the two tons of the elephant to be related to other pointer readings, i.e., to values of other variables, such as velocity, coefficient of friction, etc. Operational definitions, by providing empirical correlates for concepts in the form of variables, allow variables to be related one to another. Propositions that express relationships among variables are "scientific" in the sense that they take the form of generalizations and serve an explanatory function: if verified, they assume the character of laws; if awaiting verification, they have the status of hypotheses.

To the extent that problems of organizing and retrieving information are definitional in nature, solutions to them can be approached by introducing operational definitions. An example of a productive operational definition is the precision-recall measure, which was developed to measure the degree to which a given retrieval system does or does not achieve its discrimination and collocation objectives (Cleverdon, 1962). Precision measures the degree to which the system delivers only relevant documents and is defined as the number of relevant documents retrieved divided by the total number of documents retrieved, expressed as a ratio or percentage. Recall measures the degree to which the system delivers all relevant documents and is defined as the number of relevant documents retrieved divided by the total number of relevant documents, again expressed as a percentage or ratio. The use of these measures in quantifying the discrimination and collocation objectives makes it possible to generalize about the impact of various factors on retrieval effectiveness. One of the earliest factors studied was indexing depth, the number of index terms assigned to a document. The more index terms assigned—or, alternatively, the more access points a document admits of—the higher the recall, the lower the precision. This is, in part, the scientific explanation of why keyword searching nearly always results in infoglut.

Operational definitions are constructive; however, not all operational definitions interpret concepts as variables. Some are constructive in the procedural sense of specifying a definiendum, i.e., stipulating how the object being defined can be recognized—the conditions needed to identify it. For example, a particular kind of cake, such as a Tosca torte, might be constructively or operationally defined by its recipe.

Procedural definitions are useful when it comes to defining the ontology of a retrieval language—its entities, attributes, and relationships. Consider, for instance, the entity work, which figures in the language used to describe information-bearing documents. Conceptually a work is an abstract Platonic concept. A work consists of a certain amount of delimited information—some piece of intellectual or artistic content. Operationally, a work can be defined in terms of the procedures to be followed to construct a set of documents that contain essentially the same information. A constructive definition would specify how members of the work set can be identified, e.g., as being a transformation of a given ur-document by relationships that preserve identity, such as revision, abridgement, or translation. Without an operational definition of a work it would not be possible for a retrieval system to automatically collocate, i.e., form the set of—all documents containing essentially the same information, e.g., all editions of Dickens' *Bleak House.*

Another entity that has been procedurally defined is *subject.* Early definitions, beginning in the 1960s, were based on simple word frequencies, e.g., the more frequently a substantive word occurred in a document, the greater the probability it was indicative of what the document was about.

Over time the operational techniques used to identify subjects have become increasingly sophisticated, incorporating different types of frequency distributions, parsing, and grammatical analyses and inferences based on similarity algorithms. In addition to their usefulness in automatic indexing, these techniques have the potential to improve indexing consistency and are a *sine qua non* for the automatic merging and translating of knowledge representations.

Useful as they may be, nevertheless questions can be raised about operational definitions: (i) how valid are they, i.e., how well do they define what they intend to define, e.g., concepts such as precision, recall, work, and subject? and (ii) How valid are the knowledge representations that depend on operational definitions, i.e., how expressive are they in their representation of knowledge? Insofar as they introduce quantification, operational definitions are subject to the charge that they oversimplify. For instance, the precision and recall measures have been faulted for oversimplifying the subjective concept of relevance. Automatic techniques for identifying subjects of documents have been faulted for being term- rather than concept-based. All operational definitions lack validity, to some degrees but this does mean they are ineffective. For instance, to improve keyword searching, operational procedures used to automate the assignment of descriptors to documents are advantageously being adapted for use in online search engines. In the pursuit of knowledge, oversimplification and abstraction can sometimes be valuable in clearing away confusing linguistic underbrush to get at a clear picture of a phenomenon: outstanding examples are the precision and recall measures that developed half a century ago and continue to prove productive in advancing our understanding of the factors that contribute to retrieval effectiveness.

The Referential or Picture Theory of Meaning

The referential or picture theory of meaning also derives from an empiricist view of knowledge. This theory is consistent with, but less radical than, that of the logical positivists in that it does not demand verifiability. Its chief (and most brilliant) exponent was Ludwig Wittgenstein in his *Tractatus* (Wittgenstein, 1961/1921). The basic tenet of the picture theory is that the extensional meaning of a word is its referent. For example, the extensional meaning of the word "butterfly" is the set of all past, present, and future butterflies. Words whose referents are things in the real world can be taught by ostensive definition, simply by pointing to their referents. A child learns the meaning of "butterfly" when someone points to a butterfly and says the word. For words that have no ostensive reality, referents are postulated in the form of concepts, e.g., the referent of the general noun "beauty" is the concept of beauty. Words are contained in propositions, and these propositions, deriving directly or indirectly from sensory experience, express properties and relationships. Empirical propositions picture reality. A proposition has empirical meaning if and only if it corresponds to (pictures) reality. "Snow is white" is a true proposition if and only if snow is white. Matters of fact or states of affairs are expressed in true propositions—those that represent real knowledge about the world, e.g., "Snow is white," "Tosca was written by Puccini," and "Afghanistan borders on China." Such propositions contain words whose meanings are relatively fixed and can be formulated in first-order predicate calculus. Knowledge consists of the totality of true propositions—the totality of accurate pictures of the world.

In addition to propositions that picture the world, the empiricists recognized those that are tautological, in the sense of expressing logical relationships among propositions. It was David Hume who first distinguished the two types of propositions: those that express matters of fact and those that express relationships among ideas. An example of a tautological relationship is the equivalence relationship, e.g., "Bachelors are unmarried men." Another is the logical relation of inclusion, e.g., "All parrots are birds." This relationship is logical in that it forms the basis of deductive reasoning as exemplified by the classical syllogism: "All parrots are birds. Polly is a parrot. Therefore, Polly is a bird." Logical hierarchy employing the inclusion relationship is used in classical approaches to definition whereby a general noun, regarded as a class, is defined first by its genus and then by the characteristics or differentia that distinguish its members from those belonging to other subclasses of the genus. The knowledge hierarchies resulting from definitions constructed in this manner—for instance, the biological taxonomies—represent descriptive knowledge of essences and as such are seen to mirror the formal structure of external reality.

The picture theory of meaning, and its corollaries, particularly the one that holds that true propositions can be formulated within a logical calculus, has been one of the great generative conceptions of the twentieth century. It still holds sway. In the bibliographic area, it inspired Feibleman to develop a theory of integrative levels, in which the order of classes in a classification reflects reality conceived as a hierarchy of organized wholes (Feibleman, 1954). Ranganathan built his classification on the analogy with a meccano set, assuming thereby that all knowledge could be built out of a standard set of concepts and relations among them (Ranganathan, 1965, p. 20). Fairthorne (1961) promulgated the notion of mathematics of classification. He thought that a classification of knowledge could be formulated as a lattice, a form of Boolean algebra. Wojciechowski (1971, pp. 17–18) speculated about the survival value of a classification, concluding that it was proportional to the degree to which it was formalized and urged that classifications be mathematized.

Designers of information retrieval (IR) thesauri seized on the distinction implicit in the picture theory between tautological and empirical knowledge. They favored the former. Bernier, in the 1960s, argued that the relationships among terms in a thesaurus should be permanent, rather than transient, a priori rather than a posteriori, true in all possible worlds, rather than contingently true (Bernier, 1968). In other words, only tautological, in the extended sense of definitional, relationships should be expressed in the independent semantics of a retrieval language; contingent relationships should be expressed by its syntax, or not at all. The relationship between parrots and birds belongs in a thesaurus (all parrots are birds), but not that between parrots and pets (only some parrots are pets). This distinction, which has found its way into several thesaurus standards, is sometimes expressed as the distinction between paradigmatic[1] or context-free relationships and relationships that are syntagmatic or context-dependent. The distinction is important. First, it operationally defines what it means to say a knowledge representation exists above or independent of any given database. Second, the two types of relationships, paradigmatic and syntagmatic, have different roles to play in retrieval.

The picture theory of meaning lay ready at hand when computers and the discipline of Artificial Intelligence (AI) came on the scene. AI researchers wanted to develop computer programs to process information—to understand language and to model inductive and deductive reasoning. For the most part, the data structures they used to represent knowledge were founded upon a referential theory of meaning; necessarily they were also limited to relatively small sublanguages or microworlds. Winograd's natural language understanding program dealt with the discourse that could take place in the microworld of movable blocks (Winograd, 1972). For the program to work, all knowledge about movable blocks had to be assembled and represented propositionally. Implicitly it was assumed that language understanding could be reduced to the mechanistic manipulation of elements within closed data structures.

Based on a similarly reductionist assumption is Minsky's concept of frame (Minsky, 1974). A frame is a network of nodes and relations. Frames are used to represent knowledge about everyday or stereotypical situations, e.g., a birthday party. Knowledge about birthday parties is of two kinds: that which is always true, i.e., true of birthday parties in general, and that which is true of only a particular instance of a birthday party. Always true or essential (i.e., definitional) properties, like always true relationships, are context-free and, thus, via inheritance or by hierarchical force, apply to particular instances.

Attempts have been made to use frame-based systems for machine-assisted indexing in the medical field (Humphrey & Miller, 1987). For example, assume a document is assigned the term "bone neoplasm." Bone neoplasm is an instance of "neoplasm by site," a term that has associated with its two attributes: histologic type and disease process. By hierarchical force, bone neoplasm can be characterized by histologic type and disease process. The indexer, thus, is prompted to supply values for these attributes.

To evaluate the effectiveness and expressiveness of knowledge representations that view meaning as referential in nature, it is useful to compare this theory of meaning with another theory, the instrumental theory of meaning. This will be done in the next section.

Instrumental Theory of Meaning

Wittgenstein ended his *Tractatus* with the now famous words "whereof one cannot speak, thereof one must be silent." (One of the questions contributing to speechlessness was whether it is a true proposition that true propositions mirror reality.) For many years Wittgenstein was silent, and when he began writing again he did an about-face. He rejected the elaborate edifice of meaning that he constructed in the *Tractatus* and, in his new work, *Philosophical Investigations,* and advanced a diametrically different theory (Wittgenstein, 1953). (Such an about-face may be unique in the history of philosophy.) Its premise was that instead of defining the meaning of a word in terms of its referent, it was to be defined in terms of its use. Meaning as use. Like most theories, this one has its antecedents. Frege (Dummett, 1967) is credited with the dictum that words do not have meaning in isolation, but only in the context of a sentence. Wittgenstein spun out the implications of this dictum in almost excruciating detail.

The basic premise of the instrumental theory implies that we know what a word means when we know how to use it. Words convey meaning both in themselves and by virtue of the contexts in which they appear. Thus, what a word conveys is, in part, variable, enabling it, chameleon-like, to assume different meanings in different contexts; and, in part, fixed. The fixed part is its dictionary or context-free meaning. Some words have meanings that are more variable, more colored by context, than others. Words used in scientific discourse, e.g., mass, are for the most part fixed; their meanings are not negotiable. To change the meaning of a keyword in scientific discourse and to have this accepted would be cataclysmic; it would amount to a major paradigm shift. On the other hand, words used in, say, the social sciences are regularly used with changeable meanings, e.g., the terms "democracy" and "culture." It's almost as if changing the meaning of operative words is needed to provide a new viewpoint, thus, advancing the frontiers of sociological knowledge.

The instrumental or contextual theory of meaning if pushed to its limit would construe every word, if not as a homonym, at least as a polyseme. Polysemy is when a given string of characters has a set of different but related meanings. Homonymy is when the several meanings attaching to a character string are unrelated. Culture with its more than 100 related meanings is a polyseme. Mercury (the planet, the metal, the Greek God, the car) is a homonym. The linguistic implication of even a moderate instrumentalism is that there are more words with multiple meanings in a natural language than were ever dreamt of in the philosophies of empiricism. The implication for the design of retrieval languages is that disambiguation is a serious and very large problem. It is the homonym problem writ large, writ in the extended sense of including polysemy and contextual meaning, that is the chief cause of precision failures—i.e., infoglut—in retrieval.

The solipsistic implications of extreme instrumentalism are toyed with by Iris Murdoch in her novel *Under the Net,* the net being a semantic net. In a normal world, however, solipsism is avoided by virtue of the fact that words do have some shared public meaning. Intuitively, through experience and/or some wiring in the brain, we know how to use words. We know rules that govern their use. These rules are embedded in what Wittgenstein calls language games. Subscribing to the concept of language games entails subscribing as well to the position that knowledge representations are not descriptive of things and relations in the real world; rather they are descriptive of linguistic behavior. The use of knowledge representations to organize information is one kind of language game, one kind of linguistic behavior.

To lay the groundwork for a discussion of the implications of the instrumental theory of meaning in the design of knowledge representations for IR, it is useful to begin with why the picture theory of meaning was found wanting.

- First, the picture theory assumes a universal form of language in which the meaning of propositions picturing the world are prescribed, relatively fixed, and generally understood. The objection here is that pictures can be differently interpreted. A cup is half full or half empty. A picture of a duck from another viewpoint could be a picture of a rabbit; a picture of a block could be interpreted as a triangular prism.
- Secondly, the picture theory implies fixity of reference. But the meanings of words are not necessarily fixed in the sense of referring to sets of homogeneous objects in the real world or clearly delineated mental concepts. Many words have fluid boundaries. (A chair with three legs is still a chair.) Fluidity is necessary if words are to function in a variety of different contexts. The picture theory falls down particularly in the case of abstract words whose referents are mental constructs[2] and function words, such as adverbial particles and prepositions.
- A third problem with the picture theory is that it represents knowledge of the world as the conjunction of knowledge of independent microworlds. To regard the totality of knowledge as a simple aggregation is simplistic. Winograd's block world and Minsky's birthday world have been criticized on the grounds that it is not possible to isolate microworlds (Dreyfus, 1981). Knowledge is an encyclopedic tangle of interrelated propositions. It is all of a piece; it cannot be fragmented. Not surprisingly, a critical goal of AI research today is to develop an encyclopedic representation of knowledge. An example is the research being pursued by Lenant and his team (*Los Angeles Times,* June 21, 2001, pp. A1, A18). Their Cyc database consists of over a million propositions,

> but in addition to this it contains information about the use of hundreds of thousands of root words, names, descriptions, and abstract concepts. For instance, a Cyc robot knows that anthrax can mean the heavy metal band, a bacterium, or a disease. More significantly it knows that while a piece of wood can be broken into smaller pieces, a table cannot be broken into smaller tables. Knowing rules for the use of words, it "understands" language behavior.

In sum: the picture theory lacks expressive adequacy; it does not adequately represent knowledge. Knowledge is elusive, dynamic, and kaleidoscopic. One cannot take a snapshot of it—or, as Heraclitus observed, one cannot step in the same river twice. Wittgenstein in the *Investigations* dealt a death blow to traditionalist empiricism and the idea that knowledge was reducible to sense perceptions embodied in elementary propositions. Language, when it is released from its picturing role, is free to go on a holiday. (However, too much holidaying is where madness lies.) Thus, instead of speaking about how we know reality, we talk about the different conceptual schemes we impose on reality. Wittgenstein likens a concept to a style of painting and asks, Can we choose one at pleasure?

What then are the implications of the instrumental theory for the design of knowledge representations? They are profound and include

- what we understand by classes, e.g., the classes in a taxonomy or classification;
- what we mean by subject;
- how we design relationships in a knowledge representation, such as a classification, thesaurus, or topic map;
- how we disambiguate terms to improve the precision of retrieval;
- how we go about trying to achieve semantic interoperability, solving the problems involved in the merging of knowledge representations and in creating a universal representation from a set of representations with specialized domains.

The traditional way of looking at categories or classes is tied to an objectivist theory of knowledge whereby a knowledge classification mirrors reality. The backbone of traditional classes is the logical genus-species relationship. The guiding rule in such classifications, first stipulated by Aristotle, is that classes should be mutually exclusive and totally exhaustive—the reason being there should be no cross-classification in nature. Membership in a given class is defined in terms of essential properties; that is, two members belong to a class if they share the same essential property(ies) to a sufficient degree. Essences or properties can be expressed as specifications or conditions for class membership. It is this view of class formation that is exploited in automatic procedures used to identify entities such as works and subjects.

But are there common essences? Wittgenstein questions the tendency to look for properties common to all entities subsumed under a common noun. When the meaning of a word does not describe reality but is a function of language behavior, the instances of its use need not all share a common essence; they can be similar to one another in different ways. The categories represented by general nouns are similar to families: their members belong there by virtue of sharing family resemblances. Some have the same nose, some the same eyes, some the same tail; there need not be a single property that all share.

The idea that categories are formed on analogy with family resemblances, rather than by matching on a given property, has been another of the great generative ideas of the twentieth century. It underlies the paradigm shift in the biological sciences that led to the adoption of the methods of numerical taxonomy, which has had spectacular success in challenging the traditional biological dichotomy that divides beings into those that are organic and those that are nonorganic. In the fields of logic and computer science, it is manifested in fuzzy set theory and in bibliographic classification by the use of ambiguity operators. In IR operational definitions, using similarity matrices of family-resemblance-type categories has advanced techniques of automatic classification. An example is the U.S. Census (PACE) system, an expert coding system developed to analyze U.S. census response forms (Creecy et al., 1992). Employing a vocabulary consisting of 800 industry and occupation categories, the system assigns terms to a candidate response form by comparing it with other forms that have been manually indexed—a large number of them. At a rate of 10 responses per second, the system was able to classify 22 million responses in three months—a task that if done manually would have cost 15 million dollars in labor costs. The system was reported to perform with an accuracy rate of 86 percent.

The instrumental theory is relevant to how subjects are defined. In challenging the limits of propositional knowledge, it implicitly challenges the traditional view of what is meant by a subject. This traditional view is based on a grammatical model implicit in positivistic approaches to meaning (Svenonius, 1994). In grammar, the subject of a proposition denotes the object spoken of, which could be a concept or a thing in the real world. The role of the predicate of the proposition is to say something about that object. The proposition "Snow is white" has as its subject snow; its predicate says something about snow, *viz.,* it is white. A number of propositions about snow collected into a document would warrant saying that the document is about snow or has as its subject snow. (That there is a repeated mention of snow in these propositions is the rationale behind frequency-based techniques of automatic indexing.) Extend the model further: a sufficient number of documents about snow collected into a systematized body of knowledge about snow would warrant hypostasizing a subject named Snow.

The traditional view of what a subject is belongs to a reductionist, positivistic theory of knowledge. As such it is simplistic. Subjects are complex and at times linguistically indeterminate. They are complex to the extent that they represent not a single concept, but a system of concepts. As noted, instrumentalism holds that the meanings of words—and, thus, words used to name subjects—are in part fixed and, in part, variable. The variable part assumes its value by being contextualized within a system of concepts. Any use of a word or words to name a subject emphasizes one of these concepts more than others. "Basil" to a gardener has different connotations than "basil" to a cook. Polysemy abounds.

Some subjects cannot be named; they are linguistically indeterminate. Susan Langer in her book, *Philosophy in a New Key,* introduced the idea of different kinds of symbolism (Langer, 1949). Music and art, on the one hand, and written text, on the other, represent different symbolic transformations of experience. Only the latter, which employs a discursive symbolism, is capable of conveying propositions. Sometimes music and, more often, art are representative in that they are about something: they have a subject. Richard Strauss believed he could make music depict the cups, plates, and silverware on a table. Paintings in medieval churches depicting the lives of the Holy Family, apostles, and saints were used to enlighten those who could not read. Generally, however, what is expressed by music and art uses a presentational symbolism and tends to be linguistically indeterminate in the sense that its subject, if it has one, cannot be encapsulated in a word or phrase. Linguistic indeterminacy, however, is not limited to art and music. Poetry often uses a presentational symbolism; belle lettres can have subjects it would take an extended exegesis to depict. It may be possible to describe in an essay what *Moby Dick* is about, but this cannot be named.

Nor are the various relationships employed in IR nameable. Traditional subject-heading lists, thesauri, and classifications use generic see-also and related-term relationships to link subjects. Attempts have been made to regularize these relationships, that is, to introduce some consistency into their assignment by stipulating the conditions under which they can be used, e.g., the related-term relationship can be used between two terms if they are associated by cause-effect, symptom-disease, industry-product, etc. However, even a cursory analysis of related-term relationships in thesauri will show many to be un-nameable.

An interesting, if somewhat quixotic, attempt to cope with un-nameable relations was made by Farradane (1970).[3] He argued that Boolean operators were too generic to provide useful linking information and that the specific relationships needed for discrimination could be derived from a study of cognition. To this end he developed a system of relational operators based on Guilford's psychology. The system involved two mechanisms: three stages toward complete association and three stages toward complete distinction. The combination of these yielded nine categories of relationships, the names of which are somewhat arbitrary.

Attempts to show the effectiveness of these relationships in retrieval

Table 1.

	Awareness	Temporary Association	Fixed Association
Concurent	/ø concurrence	/* self-activity	/; association
Not distinct	/= equivalence	/+ dimensional	/(appurtenance
Distinct	/) distinctness	/– action	/: functional dependence

have not been conclusive. This does not mean cognitive psychology, which has become considerably more sophisticated since Farrandane's time, might not yield useful relationships. It seems more likely, however, that useful relationships can be discovered by studying linguistic behavior as evidenced in users' online search and navigation maneuvers.

The instrumental theory impacts how we disambiguate terms to achieve precision in retrieval. To some extent the resolution of homonyms and polysemes can be effected in the vocabulary of a retrieval language by the use of parenthetical qualifiers. But given the meaning-is-use philosophy, wherein most words are to a degree polysematic, it cannot all be done there. Some of the burden must be shouldered by the relational semantics of the retrieval language as these are lodged in its vocabulary structures (e.g., subject-heading lists, thesauri, and classifications), in its syntax or in both. Retrieval languages differ with respect to the types of relationships they express and where these relationships are articulated. An example of disambiguation using vocabulary structures is the placement of "mercury" in a number of different hierarchies, e.g., Greek mythology, metals, etc. An example of disambiguation using the syntax of a language is the placement of "flight" in the syntagm "flight of stairs."

Knowledge Representations Embodied in Thesauri and Classifications

As was mentioned in the beginning of this paper, the problem of knowledge is approachable through language analysis and theories of meaning. The theory (or theories) of meaning subscribed to by a retrieval language entails a particular way of representing knowledge. The various knowledge representations embodied in retrieval languages can be evaluated with respect to their validity, their effectiveness in achieving objectives, and their amenability to automation and semantic interoperability. To illustrate this, it is instructive to compare two different types of retrieval language, thesauri and analytico-synthetic classifications, e.g., the Dewey Decimal Classification (DDC).

Two structural differences distinguish retrieval languages that use thesarui from those using analytico-synthetic classifications (Svenonius, 2000, chapters 9, 10).[4] The first difference is in their relational semantics. Many thesauri, in accordance with established standards, try to limit their tree structures to paradigmatic relationships, whereas classifications such as the DDC freely admit of both paradigmatic and syntagmatic relationships. As noted earlier, paradigmatic relationships are those that are context-free, definitional, and true in all possible worlds. Syntagmatic relationships are space-time dependent, a posteriori, empirical, synthetic, and often transient. Another way of drawing the distinction is between relationships that are logical, e.g., the all-some relationships used in classical syllogisms; and those that are psychological in nature in the sense that they reflect ordinary language behavior. The former may be said to mirror the world as seen by the logical-positivistic lens; the latter creates the world as this is seen through the use of language.

Chief among the syntagmatic relationships embodied in classificatory tree structures are the perspective hierarchies. These hierarchies serve not so much the function of defining scientifically—though they do this to some extent—as to indicate a point of view or method of treatment. Whereas in a thesaurus, following the all-some rule, rats would be given only the broader term “rodents,” in a classification; where this rule is relaxed, rats also might be hierarchically related to laboratory animals. Moreover, in a traditional thesaurus, the kinds of relationships that abstract concepts participate in are limited, if for no other reason than they have fuzzy boundaries. How, for instance, is a term such as “freedom” to be treated in thesaurus construction? To determine if it has a broader term, one must ask, All Freedom is ———? Possibly the blank could be filled in by “liberation,” but then “liberation” would have to have at least one other subclass in addition to “freedom.” Compare this with the handling of “freedom” in a classification like DDC, which has a number of different hierarchical placements for Freedom, e.g.,

300	SOCIAL SCIENCES
320	POLITICAL SCIENCE
323	CIVIL & POLITICAL RIGHTS
323.4	SPECIFIC CIVIL RIGHTS
323.44	FREEDOM OF SPEECH

Insofar as hierarchical perspectives such as these are based on literary warrant, they are reflective of linguistic behavior.

A second way thesaurus-based retrieval languages differ from analytico-synthetic classification is in their syntax. The syntax rules of a retrieval language are used on the syntagmatic axis to combine terms to form syntagms. In natural languages, syntagms may take the form of sentences; in retrieval languages they may be called statements, subject headings, strings, or chains. An example of a DDC syntagms is

323.4430976 [Free Speech in the South Central U. S.]

The contextualizing of a term in syntagm is an instance of its use; it is therefore a method of disambiguation.

While there are exceptions, e.g., PRECIS, most thesaurus-based retrieval languages index using descriptors; they do not employ a precoordinate syntax. Thesauri for the purpose of improving IR emerged in the 1950s and were predicated on the assumption that all the complicated syntax rules in subject heading and classificatory languages could be replaced by Boolean operators: the And operator would be used for discrimination, the OR operator for collocation. In time, as the limiting nature of these operators became apparent, proximity operators were introduced to exploit contextual relationships as these occurred in the natural language of the databases being searched. The relative effectiveness of pre- vs. postcontextualization is part of the broader question of the efficacy of vocabulary control, a question beyond the scope of the present paper. Suffice it to say that the precontextualization of terms in syntagms offers the possibility of structured displays to facilitate disambiguation through browsing (Svenonius, 1995).

As knowledge representations, thesauri are limited. First they are limited in what they can express insofar as they manage—and frequently they don't—to limit their hierarchy structures to paradigmatic relationships. Second, they are limited in assuming Boolean and proximity operators are sufficient to express synthetic relationships. Being limited in what they can express, they are limited as knowledge representations and, consequently, limited in their ability to facilitate precision in retrieval. In comparison, much can be said in favor of a classification like the DDC. Compared with other types of retrieval languages, it achieves a great deal of expressive adequacy by virtue of the distinctions it can make. Using a notational coding, it can express complex subjects better even than word-based systems. By virtue of its perspective hierarchies, it can express a great deal of relational information, more than can be expressed by traditional thesauri.

Trade-offs

Many years ago Shera introduced the term "social epistemology" to denote the study of the products of man's classificatory behavior, which he sees as being relative to both time and place (Shera, 1973). He believed each new age required a new classification of knowledge, which is to say each new age represents knowledge differently. But might we not go even further than this and suggest that knowledge representations, and the epistemologies upon which they are founded, are relative also to purpose? In the design of a retrieval language, a decision must be made as to what knowledge representation to adopt. An obvious choice would be to choose one that adequately expresses what knowledge is, i.e., the representation that ranks highest on a validity scale. Such a representation might be expected also to rank high in promoting precision and recall. With respect to expressive validity, a classification like the DDC ranks higher than traditional thesauri, being to a greater degree based on an instrumental approach to meaning—an approach that offers an alternative to a single picture of the world, one that is jammed into the procrustean structure of logical hierarchies and propositional calculi.

But validity is not the only consideration in the choice of a knowledge representation. It can be argued that as a knowledge representation becomes more expressive, its semantics become overrich, the plethora of choice it offers confusing, and the elaborate rule systems on which it is based make it expensive to apply. A retrieval language that incorporates an expressive knowledge representation if too elaborate does not lend itself to collaborative efforts in its creation and application. More serious, it does not lend itself to automation, in particular to automatic indexing and semantic interoperability, i.e., the automatic merging of two or more retrieval languages. This latter is a Herculean task, one that amounts to combining two or more language games and their concomitant rules for play, a task difficult in itself, and more difficult especially if the languages rest on different epistemological foundations causing them to differ in every aspect of their vocabulary, semantics, syntax, and pragmatics. Given these considerations and the need to deal in a timely fashion with the increasing deluge of information pressing upon us, considerations other than validity become relevant.

The knowledge representations resting upon the epistemological foundations of logical positivism in its operationalist and representational approaches to meaning are further distanced from natural language than those resting upon an instrumental approach to meaning. They are formalized to a greater degree and as such are simpler, more uniform, and relatively free from subjective interpretation. The objectivity they provide through definitional rigor is essential for automated applications in retrieval, is useful in insuring consistency in programs of distributed indexing, and is helpful in attempts to merge two or more retrieval languages.

Arguably, in the design of a retrieval language, a trade-off exists between the degree to which the language is to be formalized and the degree to which it is to be reflective of language use. As mentioned earlier, Wojciechowski hypothesized that the survival of a retrieval language depended on the degree to which it can be formalized. The truth of this hypothesis remains to be seen, but certainly it is true that a highly formalized language advances the twin goals of automation and distributed indexing. On the other hand, the greater the expressiveness of a retrieval language, in particular the greater its ability to convey the contextual and relational information needed for disambiguation, collocation, and navigation, the greater validity it has as a knowledge representation. Wittgenstein asked if we could choose a conceptual scheme at pleasure. We might now ask, Can we choose a knowledge representation for a particular purpose? Perhaps we don’t always need a valid representation, when a useful one will do.

Notes

1. The term "paradigmatic" is used in the indexing literature differently from its use in linguistics.
2. Plato was the first to address the referential problems attending abstract concepts. He suggested that the referents of these were instances in the real world, e.g., the referent of (the ideal of) "beauty" was the totality of instances of beauty.
3. Farradane is credited with being the first person to use the expression "information science."
4. This is a broad generalization, one that has exceptions, but it is true in the main.

References

Bernier, C. L. (1968). Indexing and thesauri. *Special Libraries, 59(*2), 98–103.

Bridgman, P. W. (1938). *The logic of modern physics.* New York: MacMillan.

Cleverdon, C. W. (1962). Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. *ASLIB Cranfield Research Project.* Cranfield, England: College of Aeronautics.

Creecy, R. H., Masand, B. M., Smith, S., & Waltz, D. L. (1992). Trading MIPS and memory for knowledge engineering. *Communications of the ACM, 35*(8), 48–64.

Dreyfus, H. L. (1981). From micro-worlds to knowledge representation: AI at an impasse. In J. Haugeland (Ed.), *Mind design: Philosophy, psychology, artificial intelligence* (pp. 161–204). Cambridge, MA: MIT Press.

Dummett, M. (1967). "Frege, Gottlob." In P. Edwards (Ed.), *The encyclopedia of philosophy* (v. 3, p. 228). New York: Macmillan.

Eddington, A. S. (1929). *The nature of the physical world.* New York: Macmillan.

Fairthorne, R. A. (1961). The mathematics of classification. In *Towards information retrieval* (pp. 1–10). London: Butterworths.

Farradane, J. (1970). Analysis and organization of knowledge for retrieval. *ASLIB Proceedings, 22*(12), 607–616.

Feibleman, J. K. (1954). Theory of integrative levels. *British Journal for the Philosophy of Science, 5,* 59–66.

Humphrey, S. M., & Miller, N. E. (1987). Knowledge-based indexing in the medical literature: The indexing aid project. *Journal of the American Society for Information Science, 38*(3), 184–196.

Langer, S. K. (1949). *Philosophy in a new key; a study in the symbolism of reason, rite, and art.* New York: New American Library.

Minsky, M. A. (1974). A framework for representing knowledge. (AI Lab Memo No. 306). Cambridge, MA: MIT Press. (Reprinted in J. Haugeland, Ed., *Mind design: Philosophy, psychology, artificial intelligence,* pp. 95–128, 1981, Cambridge, MA: MIT Press)

Ranganathan, S.R. (1965). The colon classification. In S. Artandi (Ed.), *Rutgers series on systems for the intellectual organization of information* (Vol. 4). New Brunswick, NJ: Graduate School of Library Service, Rutgers, The State University.

Shera, J. H. (1973). Changing concepts of classification: Philosophical and educational implications. In *Knowing books and men; Knowing computers too* (pp. 327–337). Littleton, CO: Libraries Unlimited.

Svenonius, E. (1994). Access to nonbook materials: The limits of subject indexing for visual and aural languages. *Journal of the American Society for Information Science, 45*(8), 600–606.

Svenonius, E. (1995). Precoordination or not. In R. P. Holley, D. McGarry, D. Duncan, & E. Svenonius (Eds.), *Subject indexing: Principles and practices in the 90's: Proceedings of the IFLA Satellite Meeting held in Lisbon, Portugal, 17–18 August 1993,* (pp. 231–255). Munich: K.G. Saur.

Svenonius, E. (2000). *The intellectual foundation of information organization.* Cambridge, MA: MIT Press.

Winograd, T. (1972). *Understanding natural language.* New York: Academic Press.

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). New York: Macmillan.

Wittgenstein, L. (1961). *Tractatus logico-philosophicus;* with a new edition of the translation by D. F. Pears and B. F. McGuinness and with the introduction by Bertrand Russell. London: Routledge & Kegan Paul. (Original work published 1921)

Wojciechowski, J. A. (1971). The philosophical relevance of the problem of the classification of knowledge. In J. Wojciechowski (Ed.), *Conceptual basis of the classification of knowledge: Proceedings of the Ottawa Conference on the Conceptual Basis of the Classification of Knowledge, October 1–5.* Pullach bei Munchen: Verlag Dokumentation.