The Role of Relational Structures in Indexing for the Humanities

Rebecca Green

College of Library and Information Services

University of Maryland, College Park, MD, USA

Rebecca Green is on the faculty of the College of Library and Information Services at the University of Maryland, College Park, MD, where she teaches in the knowledge organization area. Special emphases include database design and cognitive linguistics. Dr. Green was the program chair of the 4th International ISKO Conference.

Green, R. (1997). **The role of relational structures in indexing for the humanities.** *Knowledge Organization,* 24(2), 72-83. 28 refs.

ABSTRACT: The paper is divided into three parts. The first develops a framework for evaluating the indexing needs of the humanities with reference to four sets of contrasts: user (need)-oriented vs. document-oriented indexing; subject indexing vs. attribute indexing, scientific writing vs. humanistic writing; and topical relevance vs. logical relevance vs. evidential relevance vs. aesthetic relevance. The indexing needs of the humanities range broadly across these contrasts. The second part establishes the centrality of relationships to the communication of indexable matter and examines the advantages and disadvantages of means used for their expression in both natural languages and index languages. The use of a relational structure, such as a frame, is shown to represent perhaps the best available option. The last part illustrates where the use of relational structures in humanities indexing would help meet some of the needs previously identified. Although not a panacea, the adoption of frame-based indexing in the humanities might substantially improve the retrieval of its literature.

1. Indexing for the Humanities

As my title suggests, I will be looking at whether relational structures can play any significant and/or beneficial role in providing access to materials in the humanities. In addressing this question, I will look in this first part at the needs of indexing in the humanities and in a second part at the general characteristics of relational structures in indexing. In the third and final part I will examine the applicability of relational structures to indexing in the humanities by illustrating contrasting situations in which relational structures would or would not be necessary for effective retrieval.

In this first part, in trying to tease out the indexing needs of the humanities, I will refer to four sets of contrasts applicable to the indexing environment: (1) user (need)- or request-oriented indexing vs. a particular version of document-oriented indexing; (2) subject indexing vs. attribute indexing; (3) scientific writing vs. humanistic writing; and finally (4) topical relevance vs. logical relevance vs. evidential relevance vs. aesthetic relevance.

1.1 User (Need)-Oriented Indexing vs. Document-Oriented Indexing

I will illustrate the first contrast in connection with S. R. Ranganathan's (1964) pronouncement of five fundamental laws of library science, in which he included two laws that are converses of each other: "Every reader his book" and "Every book its reader" – The first of these laws – "Every reader his book" – takes a user-oriented approach, picturing retrieval as a process of matching users and their attendant needs with the sources that might help resolve those needs. Since the match is not based on mere personal preferences, but focuses on the resolution of specific knowledge gaps or problems, it seems more fitting to label this a user need-oriented approach; further, as user needs are usually made known through requests, this approach is also appropriately known as the request-oriented approach. The second of these laws – Every book its reader – takes a document-oriented approach, but not in the sense that we commonly use that phrase; instead I take this sort of document-oriented retrieval to picture retrieval as a process of finding readers who might engage with a document in a meaningful and especially in an appreciative way.

While these two approaches contrast, they are not contradictory. Indeed, within the humanities the convergence of these approaches represents the ideal. I would take the prototypical successful document use in the humanities to be one that involves the reading of "good literature". This is reading that engages the mind and/or heart, thus finding an appropriate reader for the document ("Every book its reader"); at the same time the engagement satisfies some emotional and/or cognitive need within the reader, thus finding a relevant document for the reader ("Every reader his book"). The importance of both of these approaches within the humanities context requires that humanities materials be indexed both from the perspective of the user with his or her needs and from the perspective of the document with its quest for an appropriate audience.

1.2 Subject Indexing vs. Attribute Indexing

The second contrast I will make reference to operates within the context of document-oriented indexing as that phrase is usually understood, that is, indexing that takes only the characteristics of the document into account. The contrast lies between subject indexing and what I will term attribute indexing. No doubt many others would include all of my attribute indexing within the broader scope of subject indexing, but I wish for current purposes to explicitly restrict subject indexing to topical indexing, indexing that reflects what a document or a user need is about. This leaves attribute indexing to reflect such other characteristics of documents and user needs as language, recency, author affiliation, intended audience, and so on. Although attribute indexing covers far more ground than subject indexing narrowly defined, it is subject indexing which has garnered most of the theoretical attention. This primacy of subject indexing within the overall indexing arena stems from long standing perceptions about the basic nature of indexing. For example, in an article that describes the differences between document-oriented and request-oriented indexing, Raya Fidel (1994, p. 572) quotes the definition of indexing given by Borko and Bernier (1975, p.8): "Indexing is the process of analyzing the informational content of records of knowledge and expressing the information content in the language of the indexing system." Fidel (1994, p. 573) then goes on to say, "The idea that index entries, much like abstracts, represent the contents of a document has led to the notion that indexing is actually the process of creating surrogates for documents, summarizing their contents."

The traditional connection between subject indexing and document-oriented indexing is amply illustrated by the foregoing quotations. The converse association between attribute indexing and user (need)-oriented indexing is also made. Carol Barry (1994, pp. 153-157) gives perhaps the best available list of attributes that should be considered in indexing documents and requests; she refers to them as "user-defined relevance criteria." The correlation between subject indexing and document-oriented indexing on the one hand and between attribute indexing and user (need)-oriented indexing on the other hand, are not based, however, on qualities inherent in the indexing types themselves. On the one hand, taking the viewpoint of document-oriented indexing, there is considerably more to say about a document than just its subject: full document-oriented indexing requires both subject indexing and attribute indexing. On the other hand, taking the viewpoint of subject indexing, there is no reason for restricting it to document-oriented indexing; a user-need-oriented approach to indexing will almost always require a subject indexing approach at its base, supplemented by attribute indexing.

As a rough generalization, subject indexing casts a net for documents that are topically relevant to a user's needs, while attribute indexing serves in turn as a filter to screen out documents with characteristics that render them inappropriate to the user's need. Alternatively, subject indexing tends to promote recall, while attribute indexing tends to promote precision. Both, of course, are important. Humanities indexing, like most other indexing, needs both approaches: subject indexing, to gather for possible consideration all the documents of potential relevance to the user's need, and attribute indexing, to discriminate between documents meeting general needs and those meeting more specific needs.

1.3 Scientific Writing vs. Humanistic Writing

The third contrast I will address distinguishes between the indexing of literature that imparts factual knowledge and the indexing of literature that reflects appreciation of human experience. I will refer to these two types of literature as scientific writing and humanistic writing, respectively. Subject analysis based on actual content, i.e., what is expressly discussed in a piece of writing, may well be a very appropriate method for describing the use of scientific writing in its aim to impart factual knowledge. But this is not a chief aim of humanistic writing, with its goal of communicating "a true, rich, intimate and vivid subjective awareness" (Butler, 1940, p. 280 quoted in Immroth, 1974, p. 74), a creative, imaginative, spiritual, and/or intellectual appreciation of ourselves and our relationships with other persons and with our world (Immroth, 1974, p. 73).

I should quickly add two comments: (1) I am not equating humanistic writing and humanities litera-

ture, although I expect the correspondence between the two to be strong. (2) There is no reason why a given writing may not consist of both scientific writing and humanistic writing, though probably mixed in unequal proportions. Indeed, the best writings will be those that blend factual knowledge and human reflection.

Having established by definitional fiat that it is not a chief aim of humanistic writing to impart factual knowledge, I can now take a purist's approach to what it is that humanistic writing does: this is literature with the potential to enlighten and delight, but also to challenge and infuriate; this is literature in which a relationship is established between human writer and human reader, a relationship that communicates at least from mind to mind, and at its best, from heart to heart. To the extent that the literature of the humanities consists of humanistic writing, its indexing needs to reflect this more subjective brand of "relevance." In this context literature is relevant to readers because it touches something deep within them, because it makes them think and feel as only humans are known to think and feel, because it leads readers to relate to writers as if somehow we humans can actually come to know each other through writing and reading.

Unfortunately, it is not altogether clear what indexing should try to reflect in attempting to establish these virtual connections between writers and readers. Just as research was required to determine the spate of extra-topical attributes that affect user perception of relevance for scientific writing, so too we need to investigate what attributes of humanistic writing contribute most to our sense of connecting with another human through the medium of literature as well as to our perceived depth of human response.

As an example, I have especially enjoyed wrestling with the writings of the American linguist Ronald Langacker, not only because I value his linguistic insights, but also because he allows himself to be viewed as a very human scholar. For instance, he writes (Langacker, 1987, p. 31):

> In all honesty, I would greatly prefer not having to discuss methodological issues ... Nevertheless, I feel compelled to discuss the methodological assumptions that have guided me in my work. I fully expect the ideas presented here to be attacked on methodological grounds, not (I like to think) because they lack scientific validity, but because I make very different assumptions from most linguists about the appropriate adaptation of scientific methodology to linguistic investigation.

To me this writing evidences both intellectual integrity and a rare balance between self-confidence and humility found only in the truly capable. Can we fashion our indexing to indicate absolute degree of intellectual integrity or relative degrees of self-confidence and humility? Perhaps not. But we could give indication in this case of reading level (reflecting the scholarliness of the writing) and incidence of first-person references (reflecting the human orientation of the writing), which, taken together, might reveal a little something of the human qualities I find in this writing. However, there are probably limitations on how accurately or how fully it is that subjective qualities can be reflected by objective attributes. Ultimately we may need to allow the communication of subjective qualities as an essential part of indexing for humanistic writing.

1.4 Topical Relevance vs. Logical Relevance vs. Evidential Relevance vs. Aesthetic Relevance

Topical relevance is the appropriateness of a document for a user need based on the topics of the need and of the document. In a study investigating topical relevance of religious literature, Carol Bean and I (Green and Bean, 1995) found that topical relevance is not always – indeed, is not usually – based on exact topic matching. It will not be surprising to learn that the topics of the user need and of documents relevant to that need may be hierarchically related, for example, when the topic of the document is more specific than the stated topic of the user need. This is altogether reasonable, since it is not uncommon when one has an information need not to know precisely wherein one's ignorance lies. The "anomalous state of knowledge" (Belkin, 1980) that produces the information need also precludes the user from being able to specify what will satisfy the need. Therefore, the topic of a document may well be more specific than the topic of the user need for which it is relevant. The surprising part of our findings was that lumping exact matching and hierarchically-related matching together still accounted for less than half of our topical relevance relationships. More often than not, the topics of user needs and of literature relevant to those needs were connected through nonhierarchical relationships, addressing, for example, purpose or cause-and-effect or instrumentality. The high incidence of nonhierarchical topical relevance relationships that we found may not generalize to other settings, but still we must acknowledge that topical relevance relationships extend beyond relationships of exact match and hierarchical relationship.

Logical relevance (Cooper, 1971; Wilson, 1973) is the appropriateness of a document for a user need based on the document's supplying information that,

in concert with what the user already knows and possibly with information in other documents, permits the user to reason deductively to a resolution of his or her need. This view of relevance assumes that a user's cognitive state can be captured as a knowledge base and that the information content of documents can be represented similarly as logical propositions. Logical relevance can then be gauged by a knowledge representation system's deciding whether the user's question can be answered on the basis of adding any of these logical propositions to the user's pre-existent knowledge base. While this view of relevance would seem to have only limited application – first, it presumes that the user need can be stated as a directly answerable question, and second, it presumes that humans will always be satisfied with the results of deductive logic, neither of which is universally true – it appropriately points out that documents can still be relevant if they contribute only part of the answer.

Evidential relevance (Wilson, 1973) is the appropriateness of a document for a user need based on the document's affecting the user's level of confidence in a possible resolution of that need. For example, a document is evidentially relevant if it leads the user to feel more sure, or conversely less sure, of a hypothesis the user may be testing out in his or her mind. Like logical relevance, evidential relevance finds a parallel in the reasoning processes of knowledge representation systems, in this case with the certainty factors of some inferential systems.

I am prepared to suggest only very vaguely how important each of these types of relevance may be for the humanities. First, I will assume that topical relevance is of major importance for all disciplines. The notion of incrementally building knowledge upon previous knowledge presupposes a continuing strain of topicality that binds that knowledge together. While this view of knowledge growth is more often associated with the natural sciences, there is no field in which scholarship can be ignorant of its predecessors. Second, I will assume that the notion of logical relevance has more limited usefulness in the humanities than it does, say, in the natural sciences and also that logical relevance has more limited usefulness in the humanities than does topical relevance. The questions asked in the humanities are less often of the directly answerable type, but where they are, logical relevance can play a role. Third, it appears to me that evidential relevance holds a particularly prominent place in the humanities. It is the nature of scholarly work in the humanities to address a general question and to gather all sorts of evidence bearing on that question. The evidence is subject to the individual scholar's interpretation, what he or she finds plausible. Two scholars looking at the same, often inconsistent, evidence may come to diametrically opposed conclusions as to its meaning, depending largely on their differing levels of confidence in the available sources of information.

Both topical and logical evidence are clearly most at home in the context of scientific writing, since factual knowledge is usually both topical and propositional in nature. Although evidential relevance reflects interpretation, as just noted, it, like both topical relevance and logical relevance, is more at home with scientific writing than with humanistic writing; after all, evidence is normally topical and propositional in nature, too. It is not until we get to the interpretation of the evidence that we start to wander off into the squishy realm of the idiosyncratic that is the bastion of humanistic thinking and writing.

What then is relevance in the context of humanistic writing? What makes a novel speak to one reader in a profound way, but not to another? What causes a carefully crafted but simple melody to be perceived as supernal when another not-so-very-different theme is considered banal and trite? Why do some tightly written and tightly argued essays uplift, while other, similarly difficult treatises merely frustrate? What characteristics contribute to the aesthetic, and indeed spiritual, experience that is the culmination of our interaction with the best that the literatures of the humanities have to offer? Surely there is no definitive answer to any of these questions, but just as surely, if we could answer them, we would have resolved a major dilemma in indexing for the humanities. At this point there are only two things I feel to say with assurance: first, that the desire for certain types of aesthetic experiences – whether calm and soothing, or emotionally invigorating, or warmly sensitive – is at the heart of the typical lay (i.e., non-scholarly) user need in the humanities, and, second, that we do not know very much about how to answer this need in our conceptual analysis of humanities literature.

1.5 Recapitulation of the Needs of Humanities Indexing

We now return explicitly to the concern that launched this part of the discussion in the first place: What are the indexing needs of the humanities? I would like to suggest that the indexing needs of the humanities are quite varied, because there is such variety in the several aspects of the humanities situation: (1) User needs within the humanities may be cognitive in nature (which is more the realm of the scholar) or may be emotional in nature (which is more the realm of the true end user). (2) The needs of the writer (composer, artist, etc.) to find an audience exist alongside the reader's (listener's, viewer's, etc.) need to find satisfying literature. (3) While the subject of a document may be important in establishing relevance,

other attributes of the document (and of the writer/composer/artist and of the reader/listener/viewer) are also often highly significant. (4) Literature in the humanities may exist to impart factual knowledge and/or to establish human connections between writer and reader. (5) Relevance can be based on both hierarchical and nonhierarchical topical relationships, as well as on logical and evidential connections (which may be deductive, inductive, abductive, etc., in nature). Relevance may also have an aesthetic basis.

2. Relational Structures

If the indexing needs of the humanities are varied, then the indexing required to meet those needs will probably need to be varied, too. Within the context of those varied needs, the concepts that need to be expressed are often complex, that is, composed of multiple simple components that are interrelated in some way. This variety and complexity affect both the identification of indexable concepts in the humanities and their expression. I will leave it to others to address further the issue of what needs to be expressed in humanities indexing and will myself now turn to the issue of how it might be best expressed. I wish to note at the outset that conventional indexing is generally fairly simple in its structure and may not be sufficiently complex for the needs identified. At the same time I should protect myself by acknowledging that I am not setting out to solve all the problems of how to express indexable concepts in the humanities, but will restrict myself to investigating how one particular type of indexing structure might resolve some of the needs identified. To this end I turn to examine how relationships can be expressed in both natural languages and index languages. [The expression of relationships in natural languages and index languages is explored more fully in Green (1995).]

2.1 The Role of Relationships

Our examination of how relationships can be expressed in index languages begins by establishing how important relationships are to communication and then by establishing the basic semantic nature of all relationships. In his analysis of the theoretical foundations of a cognitive view of grammar, from which I quoted earlier, Ron Langacker (1987, p. 214) states that all linguistic predications are of two types: nominal predications, which express entities, and relational predications, which express events and states. He suggests that few nominal predications are atomic: most of them express relationships, but differ from relational predications by not emphasizing the relationship. In a similar vein, the entity-relationship approach to data modeling recognizes a basic distinction between entities and relationships. This is paralleled by the duality in index languages between vocabulary and devices for expressing relationships (Rowley, 1992, p. 160). In both natural language and independently developed types of artificial languages, relationships play a key role.

On the one hand, the significance of relationships seems obvious, without need of special defense. On the other hand, index languages have commonly shortchanged the expression of relationships, based perhaps on a misunderstanding of the systematicity that structures the entire inventory of relationships we use. We have borrowed from linguistics the distinction between paradigmatic relationships – those relationships like hypernymy, synonymy, and antonymy that are built into the lexicon, i.e., that exist between words (or phrases) by virtue of their lexical meanings – and syntagmatic relationships – those relationships that come into being through syntactic combination. *Hot* and *cold* are related paradigmatically, as are *furniture* and *bed*. But *ball* and *window* are only related syntagmatically within phrases and sentences such as *The ball broke the window*. Unfortunately, these two types of relationships have been characterized in the context of index languages as 'semantic relationships' and 'syntactic relationships' respectively. This terminology suggests that only paradigmatic relationships are semantic, that syntagmatic relationships are not. But the truth of the matter is that the sets of relationships that can be expressed paradigmatically and syntagmatically cannot be distinguished on semantic grounds (Hutchins, 1975, pp. 36-37; Gardin, 1973, p. 145). In general, all paradigmatic relationships can also be expressed syntagmatically, as in *Hot means the opposite of cold*, *All beds are furniture*, or, more generally, *A is a kind of B*. If *ball* meant 'an object for breaking windows', then *ball* and *window* would be related paradigmatically, too. In some sense, it is an accident of lexicalization (i.e., which concepts have specific corresponding words) whether two words are related paradigmatically, although, of course, issues of motivation and reality enter into which concepts get lexicalized. The overall point to be made here is that syntagmatic relationships are just as semantic as paradigmatic relationships. The specific relationships may not be as stable as the typical paradigmatic relationship, but the relationship types that are usually expressed syntagmatically also communicate meaning. For example, set-subset *(numbers/integers)*, whole-part *(house/bedroom)*, and type-token *(palace/Taj Mahal)* relationships are often expressed paradigmatically. But relationship types such as figure-ground (The *tree* stood out against the *blue sky*), container-contents (He took the *toy* out of the *box*), source-destination (She walked from *home* to *work*), agent-action (The *actor sang*), and cause-effect (The *explosion* caused *great devastation*), which are of-

ten expressed syntagmatically, are altogether semantic. Note that the labels for these relationship types are expressible paradigmatically; it is only the specific instances of the types that must often rely on syntactic combination for their expression. The cause-effect relationship between cause and effect is almost tautologically a paradigmatic relationship, because that relationship is built into their meaning. If that relationship is semantic when expressed paradigmatically, it is also semantic when it is expressed only through syntagmatic means.

2.2 The Expression of Relationships

2.2.1 The Expression of Relationships in Natural Language

Lexicalization, the encoding of meaning, potentially complex, in single words or phrases, is one (although the most efficient) means used by natural language for expressing relationships. Other means used by languages throughout the world include word order, function words, and case endings. Languages differ in the use they make of these devices. English, for example, makes heavy use of word order and function words, but relatively little use of case endings. Latin, however, relied heavily on case endings, while its word order was less constrained.

Word order in largely uninflected languages like English tends to reflect the grammatical roles of noun phrases with relation to their governing verbs. English, for example, is an SVO language, meaning that the subject (S) of the verb (V) normally precedes it, and the object (O) of the verb normally follows it. Because of a number of factors that complicate the correspondence between grammatical roles and relationship roles, word order alone fails to account fully for the expression of relationships in natural language. (1) The first of these factors is verb selection. A given concept can often be expressed in multiple ways, for example, using different verbs. These verbs may cast a specific relationship role into different grammatical roles and hence into different word order positions, all the while expressing the same basic thought. For example, *buy* and *sell* present a common event from different perspectives: *buy* takes the buyer as its subject while *sell* takes the seller as its subject. Knowing that a noun phrase is the subject of a sentence expressing a commercial transaction thus does not identify its role in that relationship. (2) The number and roles of optional verb arguments present in a specific sentence may also affect the correspondence between grammatical roles, relationship roles, and word order (Fillmore, 1968, pp. 21-31). For example, the verb *open* has one mandatory argument, the thing acted upon (i.e., opened), and two optional arguments, the agent of the (opening) action and the instrument used to complete that action. When all three arguments are present, as in *John opened the door with the key*, the first argument is the agent, the second is the thing acted upon, and the third is the instrument. When the agent is missing, but the thing acted upon and the instrument remain, the instrument moves into initial position, as in *The key opened the door*. When both agent and instrument are missing, the thing acted upon moves into initial position, as in *The door opened*. (3) Voice is a third mediating factor. In the active voice, the subject of a verb will tend to be an agent, while its direct object is the thing acted upon. But in the corresponding passive clause, a sentence-initial subject will tend to be the thing acted on, with the agent named in a trailing by phrase.

English includes several classes of words – verb auxiliaries, conjunctions, and prepositions – known collectively as *function words*. These words are often said to "[express] primarily grammatical relationships" (*Merriam-Webster's collegiate dictionary*, 10th ed., s.v. 'Function word'), but like syntagmatic relationships, they also express semantic relationships. Of the subclasses of function words, prepositions are often the most important for expressing relationships. They not only signal that a relationship exists between the phrase formed and some other part of the sentence, but also express, often in concert with a verbal form, what that relationship is. For example, in *Congratulations were extended to the winner by all her competitors*, the preposition *to* marks the recipient of the congratulations, while the preposition *by* marks the agent that extended them. Core noun phrases in English – the grammatical subject and direct object of the verb – are, however, not marked prepositionally. Prepositions thus exercise greater influence in the expression of less central relationship roles, for example, roles of time or place, as when the location of one entity is expressed as being *above* or *below* some landmark or *in front of* it or *behind* it. In such a context the preposition carries almost the full weight of expressing the relationship that exists between the noun phrase that follows the preposition and some other element in the sentence.

In a number of highly inflected languages (e.g., Latin, German, Finnish, Russian), morphological case endings are affixed to noun phrases to indicate their relationship to other elements of the sentence and play the same general role that word order does in English. In contrast, morphological cases are of minor importance in a mostly uninflected language like English, where only the pronominal system makes a three-way case distinction (nominative, accusative, and genitive). Otherwise, English uses only the possessive case with regularity, but has alternate means for expressing possession.

2.2.2 *The Expression of Relationships in Indexing Languages*

An index language is often viewed as a constrained subset of its corresponding natural language: on the one hand, it borrows from the natural language's vocabulary both its basic terminology and conceptual structure; on the other hand, the meanings given to terms in the controlled vocabulary may not correspond exactly with the senses they have in their "natural" setting. Similarly, an index language may use some, but not necessarily all of the means used by natural languages for expressing relationships. Lexicalization, for example, is a pretty sure bet. To the extent that relationships are expressed in precombined phrasal units, prepositions may also play a significant role. But word order, which is often characterized in the context of sentences formed around finite verbs, is unlikely to operate in index languages in quite the same way it does in natural language. And where precombined units are not used, new means of forming and expressing relationships are needed.

Some document retrieval systems fail to accommodate explicit representation of relationships in any way; instead the identification of relationships, relationship participants, and the roles they play must be inferred. In such systems, for example, those based on Boolean logic, the evidence for such inferences is largely that of term co-occurrence as identified by the Boolean AND operation. Term co-occurrence, however, carries no guarantee that a relationship holds between the co-occurring terms. And even if a relationship does exist, there is no explicit indication what it is. False drops thus occur in searches based on Boolean ANDing: the multiple terms of the search statement may occur in a document or in its index term assignments without being bound to each other in the desired relationship or indeed without being related to each other at all. Proximity searching, which is based on the intuition that "words that appear in close proximity should have more to do with each other, lexically, syntactically, and semantically, than words appearing at a distance" (Haas and Losee, 1994, p. 619) is a special application of the Boolean AND. Here the extent of the text in which the conjoined terms may co-occur is restricted, defined either in terms of absolute numbers of characters or words, or in terms of structural elements of the document. While proximity searching is likely to yield better results than unbounded Boolean ANDing, it still fails to ensure that terms co-occurring within its search window are related to each other. Similarly, proximity searching is unable to confirm the precise nature of any relationship that may exist between the components of a search statement, especially in light of the systematic treatment of important function words (e.g., prepositions, conjunctions) as stop words.

A second strategy for expressing relationships in index languages can be seen in the historic use of links. In one of its uses, related terms were explicitly interconnected in a document's indexing through the creation of multiple entry points (i.e., links) to the document, each entry point constituting, as it were, a separate logical document (Sharp, 1965, pp. 123-125). Index terms were then assigned to links such that all the terms assigned to the same link were part of a single complex subject within the document, thus ensuring that terms searched for and retrieved by virtue of their co-occurrence within a single link were in fact related to each other. Although this use of links resolved part of the dilemma faced by Boolean AND – the identification of the participants in the relationship – the problems stemming from failure to name the relationship explicitly and to identify participants' roles in the relationship are left unresolved.

A third strategy for expressing relationships in indexing involves the use of role indicators. The role indicator approach involves associating with an index term a label that indicates the role played by the term's referent in a specific context. According to this approach, the cause-effect relationship between infection and disease would be represented by two separate descriptors, with each bearing an indication of its particular role: *Infection (Cause), Disease (Effect).* Although the indicator expresses the role played by a term in a specific context, the context itself may require inferential reconstruction. Thus the overall relationship may not be recoverable. Moreover, role indicators do not create, by themselves, the necessary association between relationship participants. If the indexing of a document includes multiple instances of a single relationship type, thus causing specific role indicators to be used more than once, it may not be possible to infer with confidence for a given relationship instance which entities participate in it.

2.3 *The Evaluation of Relational Devices*

If there are so many ways of expressing relationships, how can we judge which means are the most effective or indeed whether any means are really effective? Based on criteria developed by Norman and Rumelhart (1975, pp. 45-47) for desirable characteristics of the "primitive structures" of a knowledge representation system – which include relational structures – I present here criteria against which devices for expressing relationships can be evaluated, clustered into three groups.

1. The first cluster concerns whether relationships are expressed *systematically.*

- The invariance criterion states that relationships that have the same meaning should have the same expression.
- The continuity criterion suggests that any similarity between two relationships should be reflected in an appropriate degree of similarity between their expressions.
- The relatability criterion similarly holds that any relationships between relational constructs, e.g., hierarchical relationships, should be ascertainable from their expressions.

2. The second cluster relates to whether modes of expression are able to handle the *complexity* of relationships in an *efficient* manner.
   - According to the simultaneity criterion, since relationships exist as both wholes (the relationship as a unit) and parts (the individual arguments or entities involved in the relationship), their expression should capture these two aspects simultaneously. At the same time some means for emphasizing ("profiling") the whole or a part is also required.
   - The extensibility criterion notes that expressions of relationships should be capable of extension by embedding references to other relationships within them. The embedded relationships become entities within the incorporating relationship.
   - The unity of treatment criterion posits that the relationship system that is able to handle entities and relationships using a single set of structures and operations will be more efficient than the one that uses separate sets of structures and operations for entities and for relationships.
3. A final criterion forms its own cluster and concerns the naturalness of modes of expression to the human cognitive system.
   - The psychological validity criterion recognizes that, since the ultimate users of retrieval systems are human beings, retrieval system outputs are more likely to be beneficial if system operation mirrors, or is at least compatible with, human cognition and perception.

Natural language measures up well against the second and third clusters. Lexicalization provides efficiency by meeting the simultaneity, extensibility, and unit of treatment criteria. The encoding of relationships in lexically simple words, for example, *buyer*, is particularly apt at communicating a whole as well as its parts, while at the same time focusing on the whole or on a specific part, thus meeting the simultaneity criterion. The extensibility criterion is similarly met by the lexicalization process in that the whole of one relationship can become a part in another. For example, in *Writing a novel was something he had wanted to do all his life*, the verbal nominalization, *[his] writing a novel*, expresses both the whole of a communication event and the goal part of a desire event; the communication event is referentially embedded into the desire event. From this same example, we see how natural language can use a single structure, a noun phrase, to express either an entity and/or a relationship. As for the third cluster, the ease with which persons express relationships in their native languages will be taken as *prima facie* evidence of the psychological validity of natural language as a mode of expressing relationships.

Natural languages do not, however, express relationships systematically at the level needed for document retrieval. They often provide many ways to express more or less the same thought. Therefore, relationships having essentially the same meaning do not always have the same expression. Since natural language fails to comply with the invariance criterion, it also fails to comply with the continuity and relatability criteria, which presupposes invariant representation of meaning.

It is largely because of the seeming lack of systematicity in natural language that the development of index languages has been found necessary. In the simplifying process of developing an index language, certain basic qualities of relationship expression that occur without question in natural language have tended to disappear from index languages. Thus, before being evaluated against the criteria used for judging the expression of relationships in natural language, index languages must first be brought before a more basic judgment bar. Here three essential questions are asked: (1) Is the relationship explicitly identified? (2) Are all relationship participants specified? (3) Are participants tied to their respective roles in the relationship? Unfortunately, the three major devices used in index languages for expressing relationships do not measure up very well against these more basic criteria, as previously pointed out. What seems to be needed is a relational structure that identifies the relationship involved and also incorporates the good features of both links and roles, thereby specifying the participants in the relationship and the role each plays. A good candidate is the frame, a relational structure borrowed from frame semantics within linguistics, which got the basic notion out of the knowledge representation side of artificial intelligence.

For our purposes, frames are probably best introduced by contrasting frame semantics with case grammar, both of which developed from the work of linguist Charles Fillmore. Case grammar recognized a very small set of general roles, e.g., agent, patient/theme, instrument, source, goal, beneficiary. But, as Fillmore came to realize, the generality of the case

grammar enterprise caused it to "[fall] short of providing the detail needed for semantic description; it came more and more to seem that another independent level of role structure was needed for the semantic description of verbs in particular limited domains" (Fillmore, 1982, p. 115). After all, verbs from the same domain (e.g., *buy, sell*) may cast their arguments into different cases, thus causing the invariance criterion to be violated. We need instead some means of expressing relationships where the roles remain constant across a domain, no matter which verb is used.

Fillmore's later work addressed this need with "frame semantics," where a "frame" corresponds to "any system of concepts related in such a way that to understand any one of them [means] to understand the whole structure in which it fits; when one of the things in such a structure is introduced into a text, or into a conversation, all of the others are automatically made available" (Fillmore, 1982, p. 111).

Commercial transactions constitute one such frame. This frame refers to a situation in which two persons are active. One, the Buyer, exchanges Money for a certain piece of Merchandise; the other, the Seller, exchanges the Merchandise for the Money. A full structural description of the commercial exchange frame includes slots for Buyer, Seller, Merchandise, and Money? the arguments of the commercial transaction predicate (Fillmore, 1977, pp. 78-79).

By using a frame to bind all the arguments of a conceptual predicate into a single index structure, this approach overcomes the basic deficiencies of other modes of relational expression used in index languages: it guarantees that the arguments within the frame are interrelated; it expresses what the overall relationship is; and it ties each argument to its specific role within the relationship.

One might anticipate that the greater informativeness of frames would compromise the compactness of the case grammar approach. After all, frame structures would seem to apply only within well-defined situational contexts. Recent research (Lakoff and Johnson, 1980; Lakoff, 1987, 1993) has shown, however, that conceptual structures of concrete domains are routinely extended, via metaphor, to structure abstract domains. A reasonably comprehensive inventory of frame structures need not expand interminably.

Having passed the first evaluation hurdle, frames can now be assessed against the criteria used for natural languages. It turns out that frames are exemplary in complying with these criteria. Since frames are based on the structure of fairly well-defined conceptual systems and not on specific verbs, on the one hand, nor on very broad conceptual patterns, on the other hand, they are explicitly designed to meet the invariance criterion. Frames are also well-suited to comply with the simultaneity and relatability criteria of the systematicity cluster. The componential nature of the frame permits degree of similarity between two frame instances to be computed on the basis of common slots (where slots may be named by general roles) and common slot values. Similarly, relatability can be defined in terms of relationships between frame configurations, slot names, and/or slot values.

The complexity cluster of criteria is also addressed positively by frames. Again, the basic design of a frame – an integrated structure of slots with values – fills the basic simultaneity criterion by simultaneously representing whole (the overall frame) and parts (the frame's slots); some means of profiling or emphasizing the whole or any specific part or set of parts would also need to be added to the basic frame idea to meet this criterion fully. The extensibility and unity of treatment criteria are addressed by allowing (pointers to) frames to be slot values.

Finally we assess frames against the naturalness criterion. On the one hand, it must be acknowledged that many, perhaps most people, do not consciously perceive their worlds in terms of frame-like structures. On the other hand, cognitive science widely supposes that frame-like structures play an important role in human cognition. For example, philosopher Mark Johnson (1987, p. 29) asserts that image schemata (abstract frame structures such as PATH, BALANCE, and CONTAINER) are "*structures for organizing* our experience and comprehension" that arise out of our bodily experience. Given the contrast, it may not be possible to design a relational expression system that is natural from the perspective of both human perception and human cognition. If not, a possible compromise would be to retain that which is natural from the cognition standpoint (i.e., frames) for the system's internal representation and to add an interface to translate between the perceptually natural (whatever that is!) and the cognitively natural.

3. The Role of Relational Structures in Indexing for the Humanities

The suggestion was raised before that not only were the indexing needs of the humanities varied, but in many cases they were also complex and therefore in need of some means of relational expression. Although the complexity part of that suggestion is critical to my proposal that frame-based indexing be given serious consideration in humanities indexing, I will mostly permit its substantiation to be seen in the details of others of the papers presented at this conference. What I wish to do at this point is simply to look at several specific user needs in the humanities and to examine whether the use of frame-based indexing would be helpful in meeting those needs. This

will provide the basis for establishing how widespread the need for relational indexing is.

The first user need arises out of my own experience as a church organist. In planning prelude music I am concerned with a number of attributes of individual musical selections: What season is a selection appropriate for?; What doctrinal themes, if any, is it closely associated with?; What key is it in?; How long will it take to play? Although a number of conceptual elements are at issue here (season, theme, key, playing time), they are all attributes of a single entity (musical selection). The satisfaction of this user need is therefore a straightforward one that could be satisfied by a file management system; a relational approach to indexing would be unnecessary for addressing this need.

A second user need also comes out of the context of music: I am familiar with several compositions of J. S. Bach that are transcriptions of the music of others. Did other composers also transcribe his music? Here again several components are involved in the user need: J. S. Bach, composers other than Bach, original compositions, and compositions that are transcriptions; moreover, several relationships are involved: J. S. Bach is related to the original compositions; other composers are related to the compositions that are transcriptions; and the transcribed compositions are of J. S. Bach originals. In order to provide efficient (i.e., precise) retrieval, some means of making those relational connections are necessary. Otherwise we can almost guarantee that the retrieval output will contain literature about J. S. Bach's transcriptions of others' compositions, perhaps in greater abundance than literature about others' transcriptions of his music. In a frame-based system the relational complex might be represented in the following manner (the frames used are based in part on a frame-based thesaurus developed as part of my dissertation):

```
COMPOSITION-1
   COMPOSER [J. S. Bach]
   COMPOSITION [*]

COMPOSITION-2
   COMPOSER [NOT J. S. Bach]
   COMPOSITION [*]

MATCHING-1
   FIGURE [COMPOSITION-2:COMPOSITION]
   GROUND [COMPOSITION-1:COMPOSITION]
   RELATION [LIKE]

JOURNEY-1
   SOURCE [COMPOSITION-1]
   DESTINATION [COMPOSITION-2]
   VEHICLE [MATCHING-1]
```

These four frames communicate that music starting as a composition of J. S. Bach (COMPOSITION-1 as SOURCE of a JOURNEY) ended as a composition of someone else (COMPOSITION-2 as DESTINATION of the same JOURNEY) through the vehicle of transcription, represented here as a MATCHING frame. It is not totally clear whether the conjoined use of MATCHING and JOURNEY frames is the best way to represent the concept of transcription, but it should be clear that this user need is based on complex interrelationships that would not be dealt with systematically under other indexing scenarios.

Two other examples demonstrate the potential need for relational indexing in retrieval systems for literature. The third user need states an interest in fiction that shows major personal transformations, i.e., where a character moves from one pole of the continuum to the other (like rags-to-riches stories). Here a family of relationships are involved, each a type of transformation (poor to rich, proud to humble, hard and unfeeling to sweet and loving), but the exact relationships are irrelevant. Thus it not simply a matter of using, for example, "rags-to-riches" as an indexing term. What is sought here is not a specific relationship, but a fairly general relationship. In a frame-based system it might be represented by something like the following:

```
MATCHING-2
   FIGURE [X:STATE]
   GROUND [X:STATE]
   RELATION [UNLIKE]

MATCHING-3
   FIGURE [MATCHING-2: FIGURE[X:FIGURE]]
   GROUND [MATCHING-2: GROUND
      [X:FIGURE]]
   RELATION [LIKE]

JOURNEY-2
   SOURCE [MATCHING-2: GROUND]
   DESTINATION [MATCHING-2: FIGURE]
   VEHICLE [*]

LOCATION-1
   FIGURE [MATCHING-2: FIGURE; JOURNEY-2:
      DESTINATION]
   GROUND [MATCHING-2:GROUND; JOUR-
      NEY-2: SOURCE]
   RELATION [OUT/AWAY]
   DISTANCE [FAR]
```

The use of the variable X in the slot value for both MATCHING-2:FIGURE and MATCHING-2: GROUND is meant to indicate that the identity of the frame cannot be identified, but that the same frame type occurs in both places. MATCHING-2 expresses that the STATE slots of the two X frames are UNLIKE, while MATCHING-3 expresses that the FIGURE whose STATE is being expressed are LIKE. In that way we communicate both the sameness of the character who undergoes the personal transformation, but the difference of the states experienced. The

transformation itself is captured by the movement of the JOURNEY-2 frame; the VEHICLE slot value ('*') indicates that, although the slot is applicable, its value is unknown. Finally, the LOCATION-1 frame intensifies the UNLIKEness specified in MATCHING-2, transforming it into polar opposition. Without the use of a sophisticated system of relational expression, such a user need could not be addressed systematically.

The final user need reflects the practice some writers have of including pithy quotations from others in introducing various parts of their writing. For example, each chapter of Deitel & Deitel's C++: *How to program* starts with several quotations; the chapter on virtual functions and polymorphism, for instance, quotes Shakespeare, Oliver Wendell Holmes, and Alfred North Whitehead (Deitel & Deitel, 1994, p. 524), none of whom had virtual functions or polymorphism in mind when he wrote something that in retrospect has everything to do with virtual functions and polymorphism. I can only suppose that the Deitels are very well read and have prodigious memories, that they have friends (possibly librarians!) who fit that description, or that they have access to an excellently indexed book of quotations and have figured out imaginative ways to search it. What I do not suppose is that they have access to a retrieval system able to locate quotations that make points analogous to points they wished to make, but in a completely different domain. But I can imagine such a system, and – surprise! – it is based on relational structures. The frame's compliance with the simultaneity criterion would allow searches to rank output on the basis of similarity of frame representations. Searches based on analogical reasoning could be designed to give greater weight to query and document *slots'* being the same type than to query and document *frames'* being the same type. Some attention would also have to be given to how specific the query and document frames were. Considerable work would be required to develop an effective analogical retrieval system, but it almost surely would require sophisticated relational structures, like frames, to pull it off.

Such sophisticated systems have their disadvantages as well as their advantages. A major drawback of frame-based indexing systems is that at present they exist in theory only, so their advantages also exist in theory only. Not only are there development costs to be dealt with, but implementation costs would also tend to run high. Both the indexing and searching of a frame-based retrieval system would require more resources than either does in more conventional systems. Frame-based indexing may be a luxury we cannot afford.

Conversely, frame-based indexing may be a luxury we cannot afford to dismiss. Since the literatures of the humanities are cumulative, comprehensive retrieval systems face severe problems if retrieval is not precise. This problem will only continue to get worse. Note that the expense of frame-based indexing of scientific literature might not be justifiable, since much of this literature is likely to be out-of-date within a relatively short period; the contrasting durability of humanities literature essentially means that the higher cost of frame-based indexing could be spread out over a much longer period of time and therefore be more readily justified.

The real issues for frame-based indexing in the humanities, however, concern the value of users' time, as well as the value of meeting the types of needs that only this type of indexing can address. Let us briefly consider the three user need scenarios discussed above in which relational structures would be particularly useful (that is, the second, third, and fourth situations above, which become in this context the first, second, and third, respectively). In the first, having to do with transcriptions of J. S. Bach's music, the only issue is that of precision. A frame-based indexing system would probably not locate additional relevant literature that might be overlooked by conventional systems; it would simply be able to filter out literature that was not relevant to the user. This becomes an issue of counterbalancing the user's resources with the indexing resources. With the second and third scenarios, however, we have somewhat different situations. With the second situation, having to do with personal transformations, conventional indexing systems would probably allow for well-recognized transformations (like rags-to-riches) to be included wholesale in the index language, but if high recall over a broad range of transformation types were important, the user would be ill-served by non-frame-based systems. This would be all the more the case with the final situation, in which retrieval is based on analogical search. Here conventional index systems would not be very useful at all; indeed, because metaphor has a built-in analogical base, keyword searches on words with metaphorical senses might yield higher recall than the conventional use of a controlled vocabulary; however, a frame-based system should be able to yield the best results of all, since the attainment of higher recall need not come at the expense of acceptable precision, a cost almost surely to be paid in the full-text search scenario.

In the end, the question of whether the costs of frame-based systems are justified by the better retrieval they should provide can only be answered in the context of developing exemplary prototype systems and investigating the value attached to their retrieval output. Only then will we be able to begin examining whether frame-based systems deliver on their promises of increased precision (for most user needs

that are relational in nature) and of increased recall (for user needs that are analogical in nature or that concern broad relationship types) and whether their benefits justify their costs.

Note

1. Paper delivered at Research and Development in Electronic Access to Fiction, Multicultural Knowledge Transfer and Cultural Mediation via Networks, a research seminar sponsored by the Royal School of Librarianship, Copenhagen, Denmark, November 13, 1996.

References

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science, 45,* 149-159.

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science, 5,* 133-143.

Borko, H., & Bernier, C. L. (1978). *Indexing concepts and methods.* New York: Academic Press.

Butler, P. (1940). The research worker's approach to books "– The humanist". In Randolph, W. M. (ed.). *The acquisition and cataloging of books.* Chicago: University of Chicago Press. 270-283.

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval, 7,* 19-37.

Deitel, H. M., & Deitel, P. J. (1994). *C++: How to program.* Englewood Cliffs, N. J.: Prentice-Hall.

Fidel, R. (1994). User-centered indexing. *Journal of the American Society for Information Science, 45,* 572-576.

Fillmore, C. J. (1968). The case for case. In Bach, E. and Harms, R. T. (Eds). *Universals in linguistic theory.* New York: Holt, Rinehart and Winston. 1-88.

Fillmore, C. J. (1977). Scenes-and-frames semantics. In Zampolli, A. (ed.). *Linguistic structures processing.* Amsterdam: North-Holland. 55-81.

Fillmore, C. J. (1982). Frame semantics. In The Linguistic Society of Korea (ed.). *Linguistics in the morning calm.* Seoul: Hanshin. 111-137.

Gardin, J. C. (1973). Document analysis and linguistic theory. *Journal of Documentation, 29,* 137-168.

Green, R. (1995). The expression of conceptual syntagmatic relationships: a comparative survey. *Journal of Documentation, 51,* 315-338.

Green, R., & Bean, C. A. (1995). Topical relevance relationships. II. An exploratory study and preliminary typology. *Journal of the American Society for Information Science, 46,* 654-662.

Haas, S. W., & Losee, R. M., Jr. (1994). Looking in text windows: their size and composition. *Information Processing & Management, 30,* 619-629.

Hutchins, W. J. (1975). *Languages of indexing and classification: A linguistic study of structures and functions.* Stevenage, Eng.: P. Peregrinus.

Immroth, J. P. (1974). Humanities and its literature. *Encyclopedia of library and information science, XI,* 71-83. New York: M. Dekker.

Johnson, M. (1987). *The body in the mind: the bodily basis of meaning, imagination, and reason.* Chicago: University of Chicago Press.

Lakoff, G. (1987). *Women, fire, and dangerous things: what categories reveal about the mind.* Chicago: University of Chicago Press.

Lakoff, G. (1993). The contemporary theory of metaphor. In Ortony, A. (ed.). *Metaphor and thought* (2nd ed.). Cambridge: Cambridge University Press. 202-251.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press.

Langacker, R. W. (1987). *Foundations of cognitive grammar* (Vol. 1: *Theoretical prerequisites*). Stanford: Stanford University Press.

Merriam-Webster's collegiate dictionary (10th ed.). (1993). Springfield, Mass.: Merriam-Webster.

Norman, D. A., & Rumelhart, D. E. (1975). The active structural network. In D. E. Rumelhart, D. A. Norman, and the LNR Research Group (Eds.). *Explorations in cognition.* San Francisco: W. H. Freeman. 35-64.

Ranganathan, S. R. (1964). *The five laws of library science* (2nd ed.). Bombay: Asia Publishing House.

Rowley, J. E. (1992). *Organizing knowledge: An introduction to information retrieval* (2nd ed.). Aldershot, Hants, Eng.: Ashgate.

Sharp, J. S. (1965). *Some fundamentals of information retrieval.* London: Deutsch.

Taube, M. (1961). Notes on the use of roles and links in coordinate indexing. *American Documentation, 12,* 98-100.

Wilson, P. (1973). Situational relevance. *Information Storage and Retrieval, 9,* 457-471.

Rebecca Green, College of Library and Information Services, Hornbake Building (South Wing), Room 4105, University of Maryland, College Park, MD, USA. Phone: +1-301-405-2050; fax: +1-301-314-9145; e-mail: rgreen@umd5.umd.edu.