Rebecca Green  
College of Library and Information Services, University of Maryland

Development of a Relational Thesaurus

Abstract: Various shortcomings typically attend thesaural relationships: failure to support extended relevance relationships; lack of effort in identifying a common relational inventory across types of retrieval systems; limitation to binary relationships; inattention to relationships built into the meaning of lexical units. To counteract these failings, a preliminary inventory of relational structures underlying the ca. 1250 most frequently occurring English verbs is presented. The inventory is compact and corresponds to a combination of semantic role-based verb types, as identified by Chafe (1970), and image schemata, as identified by Johnson (1987). The nature of hierarchical relationships among relational structures within the inventory is surveyed.

1. Relationships in Thesauri: Desiderata

Thesauri have two components: a set of expressions that refer to conceptual entities—some being authorized descriptors, others being lead-in terms—and the relationships among them. The basic premise of this paper is that, since relationships are the glue that holds everything together, transforming isolated parts into coherent wholes, we need to exercise a degree of control over relationships similar to the terminological and structural control we exercise over the conceptual entities they bind together.

Relationships play various roles in thesauri. Some relationships show the conceptual equivalence between lead-in terms and their corresponding authorized descriptors (e.g., USE/USE FOR). Other relationships define the nature of the semantic link between two authorized descriptors (e.g., BT/NT). Yet other relationships are used with authorized descriptors to create representations of new and complex conceptual entities (e.g., AND, OR, NOT). The first two types of relationships serve primarily to help indexers and searchers locate the best access points; the third type of relationship enables the thesaurus to expand its range to concepts not explicitly covered by the authorized descriptors. Both functions are critical to the intelligent use of a thesaurus.

The relational system of the typical thesaurus is nearly exhausted by the examples just given. (Indeed, many thesauri leave it up to the retrieval system to supply combinatorial, e.g., Boolean, relationships.) The only other common relationship type is the nebulous RT relationship. The end result is that thesaurus typically use only one well-defined relationship between base concepts for which authorized descriptors are provided—the hierarchical, broader term/narrower term (BT/NT) relationship—encompassing all others in the ambiguous related term (RT) relationship.

This situation is wanting in several regards. First, a recent study concluded that the relationship between the topic of a user's need and the topic of a document relevant to that need is often not strictly a matching relationship, but may range across a wide variety of relationship types, both paradigmatic and syntagmatic (Green and Bean, 1995). In order to model the relevance relationship more fully in our retrieval systems, we need access to a richer array of relationship types than the limited set now in common use.

Second, our retrieval systems—document retrieval systems, information retrieval systems, database systems, knowledge-based systems—would be both more effective and more efficient if they were based on a common model of what entities need to be accounted for and what...
relationships exist between them. The more extended inventories of relationships so far used in database and knowledge-based systems tend to be ad hoc. Much effort could be saved if we had a systematic and structured inventory of relationships available for use across all types of information-retrieval systems.

Third, almost all thesaural relationships that have been proposed are binary. But some relationships are by their very nature n-ary relationships, meaning that they bind together more than two entities to form the whole. To use a common database example, Supplier <supplies> Part <to> Project is a ternary relationship; we need to keep all three entities—Supplier, Part, and Project—bound together in a single relationship, because we could not otherwise correctly associate together the specific Suppliers, Parts, and Projects that belong together. In developing a relational system, we therefore need to be mindful not to limit the expression of relationships to binary relationships.

Fourth, many conceptual entities are complex and express internal relationships. For example, the conceptual entity Buyer has built into it a complex relationship structure that incorporates an object (often referred to as Merchandise), previously possessed by a Seller, that comes into the possession of the Buyer in exchange for something of equivalent value, usually Money. The term buyer thus evokes a complex relationship, which is lexicalized by the single term. If our relational systems are to perform their roles adequately, such internal relationships need to be made explicit.

These comments shed insight into the basic features required of a relational thesaurus, i.e., a structured inventory of relationship types. First, in order to model relevance relationships adequately, the inventory must admit of many relationship types. The relational information needed for tasks other than document retrieval may expand the breadth of a general relational thesaurus even further. Second, a relational thesaurus should not set limits on the degree of relationships expressed (where the degree of a relationship is the number of entities it binds together), but should be guided by the nature of the relationships that need to be expressed. Third, a relational thesaurus needs to be as cognizant of relationships that are conceptually embedded into natural language terms as it is of relationships that are used to form new, complex concepts.

This paper reports on the preliminary results of a research effort now being undertaken, which addresses our need for a general-purpose relational thesaurus. The major outcome of the study will be the development of a relational thesaurus adhering to the principles stated. This will entail both identifying a reasonably comprehensive set of conceptual relationship types and determining how to organize those relationship types into a structured inventory. The results presented here include an inventory of the most basic relational structures, which will serve as the core of the larger thesaurus. It also demonstrates some of the organizational principles needed in that thesaurus, based on an analysis of Path/Journey relationships.

2. Methodological Issues in Developing a Relational Thesaurus

A relational thesaurus strives to implement terminological control over relationships, much as conventional thesauri set out to implement terminological control over concepts. Part of the import of this statement is that although the units over which we attempt to exercise control are conceptual in nature, we can only implement this control through the medium of language, which is the most effective means human beings have for the communication of ideas. As language is used routinely to convey relationships, we turn naturally to linguistic data to reveal the array of relationships that are important to us. (Note: Since all human languages are not equivalent, a relational thesaurus based on English data, as this one is, will not necessarily be of universal
applicability.)

English uses 8 parts of speech: nouns, pronouns, verbs, prepositions, conjunctions, adjectives, adverbs, and interjections. These parts of speech correspond closely to basic components of the entity-relationship model (see, for example, Fidel, 1987, ch. 5): nouns and pronouns normally refer to entities; verbs, prepositions, and conjunctions are used to express relationships; adjectives and adverbs convey attributes of entities and relationships, respectively. (Interjections, which are said to correspond to the irrational or pre-rational, stand outside the entity-relationship model.) Prepositions and conjunctions are closed classes, and thus express only a limited range of relationship types used in English; for example, the relationships they express are almost all binary in nature. Verbs, however, constitute an open class, thus covering a wide range of relationship types, both conceptually (they cover all subject domains) and structurally (the relationship types expressed by verbs are not artificially constrained to be of a certain degree).

On the structural level, a verb can be envisioned as the hub of the linguistic unit in which it appears, with spokes leading from it to the various entities it pulls together. In other words, a verb (or verb form) governs the clause (or phrase) in which it occurs. Using a specific verb may force the speaker to mention a specific entity, may render it impossible to mention an entity, or may leave the mention of an entity to the discretion of the speaker. For example, the class of Commercial Exchange events referred to as buying and selling employ a relation between 4 participant entities, as previously noted: a Buyer, a Seller, Merchandise, and Money. If we mention such an event using the verb sell, we are required to refer to the Merchandise, but the Seller, Buyer, and Money are optional, as exemplified by the non-acceptability of (1) (which might be informally acceptable, but only if the Merchandise is identifiable from context) and the acceptability of (2)-(4). If, however, we mention the event using the verb cost, as in (5), we cannot refer to the Seller, except as the prior possessor of the Merchandise (i.e., John's bike).

* (1) John sold to Harry for $60. (Merchandise absent)
(2) The bike was sold to Harry for $60. (Seller absent)
(3) John sold the bike for $60. (Buyer absent)
(4) John sold the bike to Harry. (Money absent)
(5) The bike cost Harry $60. (Seller absent)

Several observations are in order here. First, the relational structure of a verb helps define its syntactic behavior. When the verb sell is used, for example, the Seller tends to appear as the subject of the verb, the Merchandise as the direct object, the Seller as the object of a to-clause (a.k.a. indirect object), and the Money as the object of a for-clause. But these assignments are not invariant (for example, both the Merchandise and the Buyer can show up as the subject of the verb), although the relational structure binding Seller, Buyer, Merchandise, and Money holds constant. Moreover, the syntactic tendencies of sell differ from the syntactic tendencies of other Commercial Exchange verbs (e.g., buy, charge, cost, pay, purchase, spend), but they all refer to the same relational structure.

The situation just surveyed calls for identifying relational structures not for single verbs, but for sets of verbs that refer to a common event, process, etc. In linguistic parlance, such verbs belong to the same semantic field. The initial phase in the process of building a relational thesaurus based on verbs thus entails (1) identifying an appropriate set of verbs, (2) grouping them by semantic field, and (3) determining the basic relational structure of each semantic field.

Verb identification in this study is based on frequency analysis data from the Brown corpus, a collection of 500 texts of approximately 2000 words each, representative of American English. Francis and Kučera (1982, 465-532) give a ranked frequency listing of the English word classes (words sharing stem and part of speech are merged into a single class), accounting for those in about the 85th frequency percentile and above. The 1250-odd verb classes from this list were
chosen as the basis of the study on the grounds that they would provide a sufficiently comprehensive sample to yield both the most general relational structures and some number of more specific structures; from the two, the variety of ways in which relational structures can be related to each other can be surveyed.

The grouping of verbs into semantic fields was approached from three distinct directions. The first approach used a clustering algorithm (Iwayama and Tokunaga, 1995) to look for co-occurrence patterns among the words found in the verbs' definitions. This approach is based on the assumption that verbs in the same semantic field would tend to mention the participant within their shared relational structure, entities that would be unlikely to be mentioned in the definitions of other verbs (unless they belong to a related semantic field). While the basic premise has some merit, there are many phenomena—notably polysemy, synonymy, and inflection—that adversely affect the attempt to form semantic field clusters based on co-occurrence patterns in definitions. Simple modification of the process—the application of Porter's (1980) stemming algorithm to control for inflection—proved inadequate. Ultimately, with major modifications incorporated to control for polysemy and synonymy, the approach might prove useful, but it is unclear that the labor commitment would be repaid. The second approach used verb groups from Levin (1993), which reflect certain aspects of verbs' syntactic behavior. But, as we have seen, differences in syntactic behavior can and do occur between verbs in the same semantic field. Consequently, the approach failed to produce fruitful results. The third approach used verb groupings within the categories of Roget's thesaurus (1911). Although this was the most subjective of the approaches, it proved the best starting point.

Determining the relational structure of a set of verbs results from identifying and characterizing their shared entity participants (i.e., their arguments), as we have already done with Commercial Exchange verbs. Several types of data sources are available to aid in this task. One type of data source presents synthesized information about verb arguments; examples are the verb frame information given in WordNet, an online lexical reference system, and Levin's (1993) discussions of the semantics of verb classes. Dictionary definitions represent an intermediate type data source; they tend to make explicit mention of a verb's arguments, but not always comprehensively. A third type of data source—the ultimate source, in that the other two source types rely on it—is corpus data in which the occurrence of various arguments/participants with verbs is exemplified.

3. Inventory of Basic Relational Structures

A concern that arises in building a relational thesaurus based on the argument structure of verbs is that the number of relational structures thus identified could be rather large. Several phenomena cooperate to reduce the potential complexity of the relational thesaurus to a manageable size: (1) The verbs within a common semantic field may share relational structure, as observed with Commercial Exchange verbs. (The multiplicity of verbs provides a means for focusing attention on different aspects of the relational structure.) (2) A major component that is transferred in metaphor is relational structure; the structure of a more concrete domain is borrowed to structure our understanding of a more abstract domain. (3) The specificity of verbs ranges broadly; we have both the extreme generality of go and be and make and the considerably greater specificity of meander and continue and manufacture. But the same basic relational structure of an agent moving from one point to another is present in both go and meander; the same basic relational structure of an object and some associated state is present in both be and continue; and the same basic relational structure of an agent engaging in a process to create an object, optionally using material and/or an instrument, is present in both make and manufacture.
The convergence of these phenomena sharply reduces the number of relational structures to be accounted for.

These factors combine to urge the identification of most basic relational structures as a starting point in the actual development of the thesaurus. Figure 1 introduces such an inventory of basic relational structures. Although small in number, this set of structures appears to account for the relationships underlying nearly all the verbs under examination. These relational structures are generally of two kinds. Approximately a half dozen of the more abstract structures (State, Action, Process, Action-Process, Benefactive, Locative, and Experiential) correspond to verb types identified by Chafe (1970), under the influence of Fillmore's case grammar. Many of the other structures (Link, Path, Container, Center-Periphery, Balance, and, from under Action-Process, Compulsion, Blockage, Removal of restraint, and Counterforce) have a more concrete basis and correspond to image schemata discussed by Johnson (1987). (For simplicity's sake, I have adopted their labels.) The basic inventory thus correlates positively with previous work with the semantics of verbs and with relational structures.

4. Organizational Structure within the Relational Inventory

The organizational structure of the conventional thesaurus is largely, often exclusively, hierarchical: \textit{B}Y\textit{N} links between descriptors account for most of the structure imposed by the thesaurus. By its very nature, the relational thesaurus recognizes additional relationship types, but the hierarchical relationship continues to wield the greatest force as an organizational device in the relational thesaurus.

In the context of entity classes, hierarchy is based on the notion of attributes: a narrower entity class inherits all the attributes of any broader entity class. Relationships are more complex than entities, and it turns out that some of the links readily recognized as hierarchical relationships in the relational context require richer explanation. As the relationships between more general and more specific verbs are examined, two patterns emerge. According to pattern 1, specificity is introduced when the identity of entities participating within the relationship is constrained to any degree (the more constrained, the more specific); it is in this sense that \textit{calve} is more specific than \textit{(give) birth}. As an extension of pattern 1, specificity is introduced when attributes of entity participants or of the overall relationship are constrained; it is in this sense that \textit{meander} is more specific than \textit{go}. According to pattern 2, specificity is introduced when multiple relational structures combine to form a more complex relational structure; it is in this sense that one meaning of \textit{capture} is simultaneously more specific than both \textit{force} and \textit{enclose}.

An examination of relational structures within the Path branch of the inventory demonstrates the broader sense of hierarchical relationship needed in this context. The basic Path image schema consists of 3 components: (1) a starting point, often called the Source; (2) an ending point, often called the Goal; and (3) a sequence of points, a Route, between the source and the goal, which is the focus of the structure.

The Journey relational structure incorporates the whole of the Path structure and adds an entity participant, a Traveler who traverses the Route between Source and Goal, optionally adding a Vehicle which conveys the Traveler along the path. The traversing itself is motion along the Route, which movement is in turn a specific kind of Action. The Vehicle is a specific kind of Instrument. Instruments are normally considered an optional entity participant in Action-Processes, so here we either need to recognize that Instruments can also play a participant role in Actions or we need to recognize that motion which is aided by a Vehicle is no longer just an Action, but has become an Action-Process. Journey is thus related to Path in a complex fashion: it not only incorporates additional entity participants (Traveler, Vehicle), but also is thereby...
<table>
<thead>
<tr>
<th>RELATIONAL STRUCTURE</th>
<th>DESCRIPTION OF BASIC RELATIONAL STRUCTURE</th>
<th>REPRESENTATIVE EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>Entity has value of attribute (e.g., emotion, height, weight)</td>
<td>mourn, tower, weigh</td>
</tr>
<tr>
<td>Action</td>
<td>Entity engages in activity</td>
<td>dance, jump, run</td>
</tr>
<tr>
<td>Process</td>
<td>Change from one state (e.g., size, color, emotion) to another</td>
<td>enlarge, darken, discourage</td>
</tr>
<tr>
<td>Action-Process</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Force</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compulsion</td>
<td>Force vector acts upon entity, to move it along path, toward or away from force</td>
<td>attract, pull, press, push</td>
</tr>
<tr>
<td>Blockage</td>
<td>Entity resists force vector acting upon it Potentially restraining entity removed/disabled</td>
<td>block, interfere, restrain</td>
</tr>
<tr>
<td>Removal of restraint</td>
<td>Force vectors oppose each other</td>
<td>escape, liberate, release, rescue</td>
</tr>
<tr>
<td>Counterforce</td>
<td>Entity causes other entity to come into existence or undergo significant change</td>
<td>oppose, conflict</td>
</tr>
<tr>
<td>Production</td>
<td></td>
<td>build, form, generate, make</td>
</tr>
<tr>
<td>Benefactive</td>
<td>Entity is beneficiary of action or state</td>
<td>benefit, inherit, own</td>
</tr>
<tr>
<td>Locative</td>
<td>Entity is located relative to a landmark (e.g., above or below it, before or behind it, near to or far from it)</td>
<td>sink, confront, adjoin</td>
</tr>
<tr>
<td>Experiential</td>
<td>Entity experiences mental disposition</td>
<td></td>
</tr>
<tr>
<td>Volition</td>
<td>Entity makes choice</td>
<td>choose, intend, prefer</td>
</tr>
<tr>
<td>Cognition</td>
<td>Entity engages mental faculties</td>
<td>amaze, imagine, think</td>
</tr>
<tr>
<td>Perception</td>
<td>Entity engages perceptual faculties</td>
<td>gaze, listen, taste</td>
</tr>
<tr>
<td>Link</td>
<td>A is connected to B</td>
<td>connect, link, marry</td>
</tr>
<tr>
<td>Path</td>
<td>A sequence of points leads from A to B</td>
<td>arrive, follow, return</td>
</tr>
<tr>
<td>Journey</td>
<td>Possession of entity moves from A to B</td>
<td>go, walk, wander</td>
</tr>
<tr>
<td>Gift</td>
<td>Possession of entity moves from A to B, with compensation</td>
<td>distribute, give, accept, receive</td>
</tr>
<tr>
<td>Commercial Exchange</td>
<td>Message moves from A to B</td>
<td>pay, purchase, charge, sell</td>
</tr>
<tr>
<td>Communication</td>
<td>Knowledge moves from A to B</td>
<td>ask, tell, listen, read</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td>instruct, teach, learn, study</td>
</tr>
<tr>
<td>Container</td>
<td>Boundary separates inside from outside</td>
<td>enclose, exclude, fill</td>
</tr>
<tr>
<td>Center-Periphery</td>
<td>Focal point contrasts with surroundings</td>
<td>center, concentrate, focus</td>
</tr>
</tbody>
</table>
Balance
Comparison
Justice

Hierarchy
Subordination
Genus-species
Whole-part

Correspondence between A and B
Correspondence between action and reward/punishment

One entity is subordinate to another
One category is wholly included within another
One entity is part of another

correspond, equal, match
compare, equal, match
assess, judge
govern, rule, submit
generalize, include, specify
consist, consolidate, integrate

Fig. 1: Preliminary Inventory of Basic Relational Structures

transformed from a static image schematic pattern into a dynamic Action or Action-Process.

The Gift, Commercial Exchange, Communication, and Education relational structures in turn are dependent on Journey, in that each conveys movement of some Traveler from a Source to a Goal. In prototypical Gifts and Commercial Exchanges, a concrete entity undergoes physical movement, as when a present (Traveler) is physically handed by the Giver (Source) to the Recipient (Goal) or when Merchandise is physically transferred from Seller to Buyer as part of the Commercial Exchange. But not all Gifts and Commercial Exchanges involve such physical motion. Consider, for example, the giving of a promise or the selling of a birthright. Similarly, in Communication and Education, it is (the meaning of) a Message and Knowledge, respectively, that are conveyed from Source to Goal. Thus, the movement of the Traveler does not always involve physical motion, but may be metaphorical. Moreover, even with the physical motion of Gifts and Merchandise from Source to Goal, it may be that the real entity transferred is an abstract one, Possession. In this sense, some Goals may also become Beneficiaries. In other words, Gifts and Commercial Exchanges not only have the relational structure of Journeys, but also incorporate the relational structure of Possession. Similarly, the relational structure of Cognition is also built into Education. The upshot is that these relational structures are yet more complexly related to the Journey relational structure than Journey is to Path. First, they specify something of the nature of the Traveler. Second, the movement of the Traveler may be literal or metaphorical. Third, several broader relational structures interact in these more specific relational structures, possibly in ways that go beyond the capacity of multihierarchical inheritance to account for.

5. Summary

A relatively compact set of general relational structures form the basis of the large number of relationships encodable in English. The interconnections among the larger inventory of relationships promise to be very complex and stretch our current understanding of the nature of the hierarchical relationship. Extensive further investigation into both the identification and characterization of relational structures and the organizational structure of those relationships is called for. Better control over the expression of relationships portends higher precision of retrieval in cases where searches involve multiple entity participants that could occur in multiple relational configurations; higher recall could be achieved in cases where the relational structure itself is the object of search.
Notes

1. The purview of the paper is actually somewhat larger than relationships: it extends to all predicates, but emphasizes those with 2 or more arguments.

2. This research is funded in part by a 1995-1996 OCLC Library and Information Science Research Grant.


References


