Studies in Natural Language Processing

Sponsored by the Association for Computational Linguistics

# Relational models of the lexicon

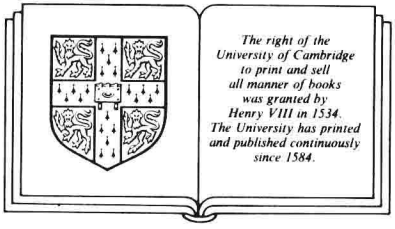
## Representing knowledge in semantic networks

Edited by Martha Walton Evens

# Relational models of the lexicon

## Representing knowledge in semantic networks

*Edited by*
MARTHA WALTON EVENS
*Department of Computer Science*
*Illinois Institute of Technology*




CAMBRIDGE UNIVERSITY PRESS
CAMBRIDGE
NEW YORK NEW ROCHELLE
MELBOURNE SYDNEY

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
32 East 57th Street, New York, NY 10022, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia



First published 1988

Printed in Great Britain at the University Press, Cambridge.

*British Library cataloguing in publication data*

Relational models of the lexicon:
representing knowledge in semantic networks
– (Studies in natural language processing).
1. Lexicography
I. Evens, Martha Walton
413'.028

*Library of Congress cataloguing in publication data*

Relational model of the lexicon: representing knowledge in semantic networks / edited by Martha Walton Evens.
p. cm. – (Studies in natural language processing)
1. Lexicology – Data processing 2. Semantics – Data processing.
I. Evens, Martha W. II. Series.
P326.R44 1988
413'.028 – dc19 88-11883
CIP

ISBN 0 521 36300 4

US

*To Leonard Evens*

## Acknowledgments

Preliminary versions of many of these papers were given at a Workshop on Relational Models held at Stanford University in conjunction with the 1984 International Conference on Computational Linguistics with financial and logistical support from the Association for Computational Linguistics. The workshop would never have happened without the help of Donald Walker and Martin Kay.

I learned about relational models from Oswald Werner and Raoul Smith at Northwestern University in company with Bonnie Litowitz, Madelyn Iris, and Judith Markowitz; all of them have contributed to this book in one way or another.

My own work on relational models would have been impossible without my colleagues, students, and friends at Illinois Institute of Technology: Thomas Ahlswede, Jeffrey Anderson, John Takao Collier, Michael Glass, Steven Gordon, Sharon King, Sun M. Li, Glenn Mayer, James Neises, Sumali Pin-Ngern, Kay Rossi, Duc Bui Tran, and Yuemei Zhang.

Sumali Pin-Ngern, Yuemei Zhang, and Sarah Evens helped patiently in the physical preparation of the text. My husband, Leonard Evens, provided essential TEXpertise at all hours of the day and night. My own department and the Mathematics Department at Northwestern University let me use laser printers and copiers at extraordinary length.

Finally, I am grateful for material assistance in editing this volume from grant IST-85-10069 to Illinois Institute of Technology from the Information Science Division of the National Science Foundation.

# Contents

# 1
# Introduction

The lexicon has become a center of attention for those involved in all problems of language. Linguists have discovered that complete analyses of both syntax and semantics require a model of the lexicon. Anthropologists cannot describe a culture without talking about the vocabulary used by the participants in the activities of that culture. Psychologists examining the development and use of language have decided that the development and organization of the lexicon is an essential part of the picture. Computer scientists have discovered that large lexicons are a prerequisite for building computer systems that interact gracefully with human beings.

Linguists, who for a long time equated linguistics with syntax and viewed the lexicon as merely a convenient storage place for exceptions to syntactic rules, have finally discovered that much of the rest of the world is convinced that language resides in the lexicon and that the function of syntax is to provide a place to record lexical regularities. Two new theoretical approaches to language, Lexical Functional Grammar and Word Grammar, reflect this change and have helped to refocus attention on the lexicon.

Anthropologists are focusing on the problems of ethnography, which involves describing a cultural milieu in terms of its sublanguage. The first step is to build a lexicon for that sublanguage. In return, research in ethnography is providing us all with a methodology for eliciting lexical information from informants when written sources are not available or not complete.

Psychologists interested in the organization of memory are necessarily concerned with the organization of lexical and conceptual information, the ways in which we access this information, and the ways we use it to build cohesive discourse. They are also asking questions about the acquisition of lexical knowledge in human beings and the evolution of the ability to make definitions. Neural network research is providing more detailed models of lexical access and other natural language processes.

Computer scientists cannot build natural language processing systems large enough to handle real world problems without figuring out how to build larger and more detailed lexicons. The move from natural language front ends and text understanding systems to text generation, machine translation, and speech systems is making overwhelming demands for much more detailed lexical information about much larger vocabularies.

This book focuses on a collection of approaches to the lexicon that use relational semantics. Relational semantics is one of three major competing approaches to the study of meaning. The first approach views concepts as forming semantic fields or domains. This approach is essentially structural: a term is defined by its place in the field. Semantic fields have been useful in descriptive linguistics, but they do not provide a theoretical framework strong enough to use as a foundation for building lexicons for parsing and text generation.

The second approach includes componential and feature analyses. Here the focus is on those common features that enable items to form a domain and also on the different features that distinguish items in a domain from each other. Componential analysis works well for closely circumscribed domains like color terms, kinship terms, and personal pronouns, but it is of limited use when we are dealing with more complex lexical domains that contain words with overlapping meanings.

The third approach, the relational approach adopted in this book, accepts the existence of semantic domains, but attempts to make explicit the structural organization that is implicit in other models, and describes how the elements of a domain are related to each other. The links that connect the elements of the domain are called lexical or semantic relations. Relations between words are called *lexical relations*. Relations between concepts are called *conceptual* or *semantic relations*. Since words and concepts are inextricably intertwined the phrase *lexical semantic relations* is often used when it is unnecessary or impossible to make a distinction.

The most familiar lexical relations are *synonymy* and *antonymy*, which are often marked explicitly in dictionaries. Often *promise* and *pledge* are listed as synonyms of each other, while *hot* and *cold* are coded as antonyms. Many other lexical relations, such as *taxonomy*, *cause*, *child*, *part*, and *sequence*, appear implicitly in the dictionary, but are used more or less explicitly in discourse and in making inferences. For example, a *lion* is a (kind of) *mammal* (the concepts are taxonomically related); to *send* means to cause *to go*; a *cygnet* is a baby *swan*; a *petal* is part of a *flower*; and *Monday* is always followed by *Tuesday*. [See Evens et al. 1980 for more examples.] We can abbreviate this information with word-relation-word triplets as shown below. We can even picture these words as nodes in a network with relation names as labels on the arcs connecting them.

| | | |
|---|---|---|
| promise | SYN | pledge |
| hot | ANTI | cold |
| lion | TAX | mammal |
| to send | CAUSE | to go |
| cygnet | CHILD | swan |
| petal | PART | flower |
| Monday | SEQUENCE | Tuesday |

This book attempts to include the best work in relational models from linguistics, anthropology, psychology, and computer science. The authors include not only university professors of anthopology, computer science, information science, linguistics, psychology, and Slavic languages and literatures, but also real-world experts on database interfaces and machine translation – and one whose job is making trucks talk to their drivers.

While the authors of the papers in this volume have all chosen to use relational models, they disagree on almost every other aspect of the care and feeding of the lexicon. The psychologists, naturally, are concerned to establish the psychological reality of relations, whereas most computer scientists think that psychological reality is irrelevant; what counts is computational convenience. The anthropologist Werner considers lexical relations only as a reflection of underlying conceptual relations, while Mel'čuk, a linguist, rejects relations that are "too semantic" as not sufficiently precise. Some agree with John White that all semantic problems can and should be solved with relational models; others like John Sowa combine relations freely with other kinds of models. But the focus of the greatest disagreement is the number of relations posited. Werner claims that all knowledge can be expressed in terms of just three relations: Modification, Taxonomy, and Queuing or Sequencing. Mel'čuk has precisely fifty-three. Evens and Ahlswede use more than 100 relations for adjectives alone. Efforts to resolve these controversies have led to much new research and debate.

This book is divided into three parts. The first part examines alternative structures for the lexicon. The second part explores the place of relational models in the representation of knowledge for a variety of applications. The papers in the last part investigate the nature of relations themselves.

The papers in the first part concentrate on the structure of the lexicon. This is an absorbing research problem for linguists and cognitive psychologists. It is also a crucial practical problem for anyone trying to construct a natural language processing system. The first chapter describes Igor Mel'čuk's revolutionary ideas about dictionaries. In the second Nicoletta Calzolari talks about the structure of the lexical database for Italian that she has constructed at the University of Pisa. In the third chapter Tom Ahlswede describes methods for the organization and construction of a lex-

icon for a medical sublanguage. Part I ends with a more theoretical and philosophically oriented paper about conceptual structures and John Sowa's approach via canonical graphs.

The papers in Part II use relations in the representation of knowledge for a variety of applications. Werner uses just three relations, *modification*, *taxonomy*, and *queuing*, to represent the whole variety of cultural knowledge acquired by C-KAD, his cultural knowledge acquisition device. Then Grimes tells us how to develop a relational database to store a lexicon full of relational material. The next chapter by John White uses relations to structure a machine translation system. Edward Fox explains how to improve the performance of an information retrieval system with a relational thesaurus. Bruce Ballard's chapter rounds off Part II with a description of the relational knowledge structure used in his natural language front end.

In Part III we turn from a study of ways to use relations as tools to an exploration of the nature of relations themselves. Should they be considered as atomic and indivisible or is it useful to try to analyze them into yet smaller components? What about important properties of relations like reflexivity, symmetry, and transitivity? Judith Markowitz leads off this part with an examination of the role that relations play in category judgments. Then Madelyn Iris and her colleagues take a careful look at the part-whole relation and decide that it is really a family of four relations – a solution that explains several of the contradictions surrounding this relation. In the next chapter, Roger Chaffin and Douglas Herrmann take a totally orthogonal approach to these same problems and propose that relations are themselves composed of smaller psychological components called relation elements. Then William Frawley gives a philosophical close to the book in an examination of metascience that illuminates the roles that relations play in the organization of the scientific vocabulary.

This introduction discusses some of the issues that arise in building relational models of the lexicon. Next comes a review of current research in the development of relational lexicons and an explanation of the place of the papers in Part I within this stream of research. The following section discusses the applications of relational lexicons illustrated by the papers in Part II. The last section describes some current research into the nature of relations and the contributions made to this research by the papers in Part III.

## Issues in the design of relational models

### Lexical vs. semantic relations

Semantic relations connect concepts; lexical relations connect words. While most models use a combination of lexical and semantic relations, some people have chosen to work with just one kind. People building memory models naturally concentrate on semantic relations. People building lexicons with words and phrases as entries need lexical relations primarily. Oswald Werner (Chapter 6) is trying to build a language universal memory model. It is not surprising that his relations are semantic. John Sowa (Chapter 5) is building canonical graphs as a part of a comprehensive memory model; his relations are also primarily semantic. On the other hand Mel'čuk (Chapter 2) and Calzolari (Chapter 3) are both involved in lexicography. Naturally their models stress lexical relations, as do the lexical databases designed by Ahlswede (Chapter 4), Grimes (Chapter 7) and White (Chapter 8).

### Psychological reality vs. computational convenience

Anthropologist, linguists, and psychologists are all concerned with establishing the psychological reality of their models. Traditionally, confirmation of psychological reality comes from native speaker intuitions or from informant behavior given a variety of tasks. But this kind of evidence for psychological reality may be difficult to judge. Language behavior is so complex that models of entirely different structure can account for the same language phenomena [Morton and Bekerian 1986].

Computer scientists are split into two camps: those who are deliberately trying to model human behavior and take the question of psychological reality seriously, and those who view human psychology only as a possible source of useful algorithms and relations as a convenient structure for organizing a lexicon or accessing individual words. Edward Fox [1980, Fox et al. 1988] is ready to include any relation that can be useful in information retrieval, without reference to its ontological status. Ahlswede and Evens (Chapter 4) also add relations as it becomes convenient. The taxonomy relation is divided into three separate relations, one for nouns, one for verbs, and one for adjectives, since the machine-readable dictionary that is the source of their data uses different definition patterns for different parts of speech and the inference process that makes use of the relational information also uses different axioms for objects and predicates. Other relations are also treated in this way.

Zholkovsky and Mel'čuk [1967/1970], in a design for a system to do automatic paraphrase, discuss the possibility of adding new relations as needed but now (Chapter 2) insist that there are precisely fifty-three and that they have them all neatly categorized and listed. A conviction of the need for

psychological reality seems to have crept up on them over the years. They do allow for flexibility in other ways – their relations can be combined in a number of ways, for which they give only a few examples, and they provide for *non-standard lexical functions*, which are too specific or too limited in range to be granted full status as lexical functions, but which are available for use in applications.

Markowitz, on the other hand, has always been convinced of the importance of establishing psychological reality. The work described in Chapter 11 is based on a large number of extensive interviews of human informants. She has also performed a series of studies on the development of definitions in children based on experience as a participant-observer and on transcriptions of taped sessions.

## Discovery procedures

The commitment to psychological reality affects the methodology used to establish relations very strongly. Those who believe in the psychological reality of relations are very properly concerned with discovery procedures for determining them; those who view relations as a convenient lexical access method are content to invent them as needed.

The anthropologists Casagrande and Hale [1967] collected 800 folk definitions from a single informant in a study of dialect differences between Pima and Papago, two Uto-Aztecan languages of the American Southwest. They later realized that they had a valuable source of data for a study of semantics and began to examine the internal organization of the definitions themselves. They classified the definitions into thirteen relation categories, using consistent syntactic cues for each category. For example, all examples categorized as *antonymy* are adjectives defined by phrases of the form "not ...." Thus, low is categorized as "not high." As native speakers of English they propose a fourteenth relation, the *constituent* relation, where "X is defined as being as constituent or part of Y," which did not appear in their data. They classify "*stinger*: it stands on the end of his [the scorpion's] tail" as a spatial relation and "*horns*: cows have horns" as *exemplification*; not as *part-whole*. This rigid methodological standard helped to establish the reality of relations in their data and to provide a basis for the use of relations in ethnography.

Although John White's focus is on the application of relational models to machine translation (Chapter 8), he discusses his methodology for establishing relationships at length and insists that rigor in the collection of data is essential to produce results that will satisfy users of his system.

Smith [1981] describes a methodology for finding relations in machine-readable dictionary definitions that has been used in a number of experiments by Ahlswede [1985a], Evens et al. [1985, 1987], and Markowitz et al.

[1986], among others. Smith made an extensive study of defining formulae, those phrases that appear in many different definitions like "of or relating to ..." or "the quality or state of being...." Smith showed that these defining formulae provide reliable clues to lexical semantic relations. Thus we can use phrase counts and KWIC indices of definition texts to discover relations. These same tools give a way to judge the relative importance or at least the relative frequency of particular relations. Smith [1985] discovered, for example, that *act* is the most frequent noun in definitions of nouns in W7 and *part* is the second most frequent.

### Paradigmatic vs. syntagmatic relations

The distinction between *paradigmatic* and *syntagmatic relations* has been debated for a number of years. Syntagmatic relations connect words that co-occur frequently; they are sometimes called collocation relations. Paradigmatic relations relate words that express the same meaning (or some part of that meaning) in some other form. In linguistics paradigmatic relations can be viewed as an extension of the relationships between members of a verb paradigm. *Swim* is paradigmatically related to *swam*, *swum*, *swims*, *swimming*, *swimmer*, *swimmingly*, and *bathe*. It is syntagmatically related to *water*, *pool*, and *bathing cap*. *Star* is paradigmatically related to *starry* and *sun*, while it is syntagmatically related to *shine* and *moon*.

The terms *paradigmatic* and *syntagmatic* have also been used by psychologists to characterize responses in word association and definition tasks. Some examples of paradigmatic responses are: *duck – drake*, *send – go*, and *lion – mane*. Some syntagmatic responses are: *duck – swim*, *send – money*, and *lion – roar*.

We get information about paradigmatic and syntagmatic relationships in different ways. Paradigmatic information typically appears in standing or generic sentences, that is, sentences that are always true:

> Adult male lions have manes.
> A drake is a male duck.

Syntagmatic information, on the other hand, appears most often in occasional sentences, sentences that describe particular situations.

> The lion roared in fury at being caged.
> We saw a family of ducks swimming in the pond.

Some models, like Fahlman's NETL, emphasize paradigmatic relations; others lay much more stress on syntagmatic relations [Smith 1984].

New research on bilingualism and new approaches to machine-translation have brought forward a new kind of paradigmatic relation, the *transfer re-*

*lation.* Transfer relations relate a word or phrase in one language to a semantically equivalent word or phrase in another language. There are many situations when a whole phrase is necessary in one language to translate what is expressed by a single word in another. Thus, the psychologists and computer scientists involved with translation models tend to be enthusiastic supporters of Becker's [1975] arguments for the phrasal nature of the lexicon. White gives a number of examples of the use of transfer relations in Chapter 8 from his work on machine translation. He argues for building a relational network of terms for each language in the machine translation system and then connecting these networks wherever possible with transfer relations.

### Lumpers vs. splitters

The greatest debate is between the lumpers and the splitters – between those whose relational models contain a small number of fairly general relations and those whose models have a large number of specific relations. Generally, models with a smaller number of relations have more concept-based semantic relations such as the models of Sowa (Chapter 5) and Werner (Chapter 6). Models with a larger number of relations tend to be populated by more surface-oriented lexical relations, as in the models developed by Calzolari (Chapter 3), Ahlswede (Chapter 4), Grimes (Chapter 7) and White (Chapter 8). Some lexically-oriented models have families of relations with the same core meaning; for example, one relation for expressing noun-taxonomy, another for verb-taxonomy, another for adjective-taxonomy, etc. There are at least two reasons for this proliferation of relations. Where relations are motivated by defining formulae, the fact that nouns, verbs and adjectives are defined differently suggests different relations for different parts of speech. Alternatively, where relations are used heavily in inference making, different parts of speech require different predicate calculus axioms. For example, if A and B are nouns and A ISA B, then we need an axiom that says that $A(x) \Rightarrow B(x)$. If, on the other hand, A and B are two transitive verbs and A ISA B, we need an axiom of the form $A(x,y) \Rightarrow B(x,y)$.

At first glance it may seem that the lumpers have all the psychological reality on their side; they can certainly claim a large chunk. But if we look again, remembering the large number of relations proposed by those eminent psychologists George Miller and Philip Johnson-Laird [1976], a second glance suggests that the lumpers are allied with the mentalist camp in psychology, while the splitters are closer to the behaviorist side. Certainly,

the easiest way to garner a large collection of relations is to treat each different syntactic strategy as signalling a separate relation.

Werner's system of only three relations (supplemented by the logical operators AND, OR, and NOT) is the smallest we are aware of (Chapter 6). Mel'čuk and Zholkovsky (Chapter 2) have exactly 53. Evens lists more than twice as many in an analysis of W7 [1981]. But Raoul Smith can definitely top that. In an analysis of adjectives defined in W7 with the formula "relating to," he identified twenty-three different adjective relations (corresponding to different adjective suffixes) [1981]. The effects of making choices among these alternatives are best seen in the designs for lexicon building described in the next section.

## Constructing a relational lexicon

The process of actually building a lexicon raises not only the kinds of theoretical issues discussed in the previous section, but many immediate practical questions. The papers in the first part concentrate on the structure of the lexicon. This is an absorbing research problem for linguists and cognitive psychologists. It is also a crucial practical problem for anyone trying to construct a natural language processing system.

Computer technology has become so pervasive that almost everyone who constructs a lexicon today will use a computer as a tool, even when the goal is a printed book. This technological change implies that scholars working on lexical problems, whether they are anthropologists, commercial lexicographers, linguists, psychologists, or computer scientists are facing a common problem. Essentially all of us are, of necessity, in the process of constructing lexical databases. Two fundamental questions must be answered by everyone who starts out to construct a lexical database: What are your goals, that is, what is the database going to be used for? And where is the data going to come from?

### The uses of lexical databases

What advantages does a lexical database on a computer give us that cannot be found in a print dictionary? If we look at the ordinary person who uses the computer only for word-processing, at first the benefits appear to be rather slight, seemingly limited to the convenience of having spelling correction, definitions, and hyphenation available online [Mark Fox et al. 1980]. With the addition of lexical relationships, however, the lexical database also becomes a thesaurus; thus word lookup is also available online.

Martin Kay [1983] has proposed a dictionary server as part of his plans for the dictionary of the future. Kay is talking about making available online whatever the dictionary user asks for – spelling, pronunciation, hyphenation, regular and irregular grammatical forms, idioms, relations and

related words. The dictionary server may need a facility for synthesizing definitions; Evens et al. [1985] suggest ways of generating definitions using defining formulae.

The advantages for an advanced learner of English trying to write essays and business letters are more obvious. The new advanced learner's dictionaries have made some information explicit for the first time. They have been an important force in the construction of lexical databases. They are not, however, easy to use. It is necessary to search backwards and forwards for material, to memorize the symbols that designate fifty or one hundred verb patterns or to look them up every time you search for a new word. A lexical database can find the verb pattern information, present it in human-readable form and provide appropriate examples.

The kind of explicit information needed by advanced learners is precisely the kind of information needed by natural language processing programs. Information retrieval systems need relational thesauri to add index terms to queries. Natural language front ends and text understanding programs need verb pattern information and verb forms for parsing. Text generation systems need even more lexical data to generate coherent text [Collier et al. 1988]. Machine translation systems need lexical databases for both languages and a set of transfer relations to record bilingual correspondences.

There are already fairly good speech synthesis systems that can read text aloud [Church 1985]. But if such a system is to be able to handle a large range of text it needs a large lexicon with phonetic information.

Lexicographers need lexical databases both as a tool and a source for information in dictionary building.

Research psychologists need lexical information of at least two different kinds. People who are setting up word recognition experiments need collections of words that satisfy particular phonetic or syntactic criteria for subjects to recognize and combine in a variety of different tasks [Schreuder 1986]. Psychologists are also interested in definition material as a subject for study in itself to determine defining strategies and to infer memory models.

A lexical database encapsulates a great deal of information about the culture that created it. Anthropologists can find semantic fields and other ethnographic information organized for retrieval in a lexical database. Furthermore, a lexical database is an essential tool for making models of informants [Werner 1978].

Linguists need lexical databases to support the development of grammars both at the sentence and discourse level. But the most obvious use for lexical information is in the study of semantics by linguists and philosophers. John Olney [1968] made the first machine-readable dictionary by keypunching *Webster's Seventh Collegiate Dictionary* with a series of philosophical