DOCUMENT RESUME

ED 083 812 EM 011 533

AUTHOR	Dostert, Bozena Henisz; Thompson, Frederick B.
TITLE	Verb Semantics in a Relational Data Base System.
INSTITUTION	California Inst. of Tech., Pasadena.
SPONS AGENCY	National Science Foundation, Washington, D.C.; Office of Naval Research, Washington, D.C.; Rome Air Development Center, Griffiss AFB, N.Y.
REPORT NO	CIT-REL-R-6
PUB DATE	Jan 73
NOTE	25p.; See also EM 011 530 through EM 011 532 and EM 011 534 through EM 011 536
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Computational Linguistics; Computer Programs; Computers; *Computer Science; Program Descriptions; Programing; *Programing Languages; Semantics; Sentences; *Verbs
IDENTIFIERS	Meaning; Predicates; *Rapidly Extensible Language; REL

ABSTRACT

Ways of making the natural language of unsophisticated computer users meaningful to the computer are discussed. The discussion is set within the context of the Rapidly Extensible Language (REL) System, a question answering system with underlying relational data bases. Major topics covered include individuals and predicates, the problem of verbs, and verbs in REL English. The essential points stressed are that the meaning of a sentence depends upon the contents of the data base to which it refers, that the tying string of a sentence is its verb, and that the sentence patterns take on meaning as they fit within the broader fabrics supplied by context and reality. Examination of examples shows that verbs radically shift their semantic content with shifts in the context in which they are used, and it was concluded that it was doubtful that there was any single method analyzing the complex webs that verbs set up among their assiciated nouns in any sentence. (Author/LB)

a project report on

REL

Co-principal investigators:
Bozena Henisz Dostert
Frederick B. Thompson

California Institute of Technology
Pasadena, California, 91109

Verb Semantics in a Relational Data Base System

Bozena Henisz Dostert

Frederick B. Thompson

January 1973

REL Project Report No. 6

This research has been supported by:
(1) Office of Naval Research contract #N00014-67-A-0094-0024
(2) National Science Foundation grant #GJ-31573
(3) Rome Air Development Center contract #F30602-72-C-0249

California Institute of Technology
Pasadena, California 91109

Note:

A preliminary version of this paper was presented at the Third Annual Meeting of the Association for Computational Linguistics at Chapel Hill, North Carolina, July, 1972, under the title "Verb Semantics in REL Case Grammar".

A thoroughly revised and updated version of this paper was presented at the ONR Symposium on Text Processing, in Pasadena, California, November 1972.

Both versions benefited from discussions at both conferences, which provided the advantage of not requiring final written versions at the time. This written version has been prepared for inclusion in the Proceedings of the ONR Symposium.

We express our thanks to the organizers of the two conferences and to the participants in both, whose comments and discussions we found very profitable.

B. H. D.
F. B. T.

Contents

Introduction

One of the immediate tasks of computational linguistics is to make it possible for a researcher untrained in computer science to communicate with the computer in natural language. Researchers, managers and other users of data and builders of models need a natural medium for asking questions and inputing data concerning the universe of discourse that is of interest to them. They need to be able to communicate with the computer in statements and questions representative of the sentences of their natural language. Our task is to find ways to make these sentences meaningful to the computer. There are many and diverse reasons why communication with computers in natural languages is desirable, not the least of which is that it provides for an enlargement of the user community. Such an increase in our efficiency of expression is mandatory to keep pace with the availability of computers themselves.

The central point of this paper may be summed up in the paraphrase of the saying that "beauty is in the eye of the beholder," which would be that "meaning is in the data base of the user."

The environment in which our discussion is set is that of a question-answering system with underlying relational data bases called the REL (Rapidly Extensible Language) System. The system is designed to allow users to communicate with their data in languages that are natural to the given users. The notion of a language "natural" to the user needs some explanation here. Programming languages are very different from the languages of ordinary discourse. On the one hand they are formal and limited in their syntax. A more important difference is that they instruct the computer to carry out procedures rather than focus on the subject matter to which these procedures apply. Meaning in natural languages has to do with ships and shoes and sealing wax and cabbages and kings. When sentences about such things are stated in a natural language, how are they to be interpreted meaningfully by computers? So, the important question is, what is meant by "natural" languages.

The view seems fairly widespread that there should be a single, all encompassing language that spans the wide range of meanings to be found in the many uses of English. At the same time, this single language should supposedly reflect the subtle nuances of meaning that each fluent native speaker can discriminate when using English in his own specialty. A contrary view, which we will elaborate on further, holds that natural language can be delimited be definition and description only when it is limited both in its domain of discourse and in its meaning discrimination. In this view, natural language is essentially unbounded in its ability to express varieties and shifts of meaning. And this is precisely where the greatest expressive power of natural language lies. This becomes especially evident when the speaker himself increases his own understanding of the area of his investigation, which is, obviously, the prime area of his discourse, and thereby extends the very context in which he is operating. Context. as is well known, goes far in determining meaning. It may indeed be that "natural" language is always relative to context and that the narrowness of the limits of the domain of a discourse goes hand in hand with the subtlety of meanings which that discourse can sustain. It may very well be that the all encompassing English of the mythical fluent native speaker is not "natural" at all.

It is worth while to quote here at length from recent findings on specialized languages, especially with respect to that mysterious, and plaguing aspect of natural language, ambiguity.

> "One notable difference between the surgical jargon and other specialized languages was the situation with respect to ambiguities -- syntactic and otherwise. There were syntactic ambiguities in the surgical reports but they were not nearly as troublesome as the ambiguities in other narratives. For example, one of the most common sentences in the surgical reports was the closing remark:
> 'The patient left the operating room in good condition'
> From a syntaxtic viewpoint this might be read as a short form of: "The patient finished mopping the floor and left the operating room in good condition."
> Whenever this syntactic ambiguity was pointed out to the surgeons they were both surprised and amused, because no surgeon would read the report in this ambiguous way. Indeed there is really no semantic ambiguity here because the universe of discourse is severely restricted in this jargon. A physical description of the operating room is not something which would appear in a surgical

> report. To put it another way, there are rigid conditions that determine the admissibility or acceptability of a sentence in a surgical report -- conditions peculiar to this reporting jargon -- and the syntactic ambiguity is resolved by these restrictions. There is therefore no ambiguity in the actual transmission of information." [1]

When it comes to the immediate task of constructing natural English for communicating with the computer, the issue of an all encompassing language is, at least for the near term, pre-empted. This is so for two reasons. First, our understanding of the mechanisms of language and our systematic collection of linguistic evidence are at this time adequate to guide only rather gross solutions to this problem. We will say little more about this limitation, arising from the youth of our field, because the semantic power of English for the computer is far more limited by another aspect of the problem.

The language we use in communicating with the computer cannot discriminate nuances of meaning which are not discriminated in the data available to the computer. If the data available in the data base concerns family relationships in a certain human community, the question whether or not a fluent native speaker can distinguish the following three meanings of bachelor: an unmarried male, a graduate of a college, or a mateless seal is quite beside the point. At this time, and for the forseeable future the domains of discourse relevant to our computer systems will be limited, even though we may fret at being earth-bound by practical considerations of hardware and costs. The practical task we face is to fit language, data availabity, and the universe of discourse into a coherent and balanced whole.

Individuals and Predicates

We said above that the meaning of a sentence depends on the data base. We need to review the nature of data bases and algorithmic models on which the semantics of a natural language must rest. For this purpose, it is sufficient to note that data and models are both abstractly characterized and concretely realized as entities or individuals, on the one hand, and as relationships among them on the other. Thus, the United States census is a body of data which categorizes individual citizens into classes, and gives their relationships to each other and to various measurements, such as age and income. The universe of discourse of a language system pertaining to such data can be thought of as a given collection of entities or individuals and certain properties and relationships that are predicated about them.

In typical computer data systems, this abstract characterization is concretely represented in a network in which each individual is associated with a record whose successive entries are the values for this individual of certain relationships. For example, the record for Tom Jones may read:

male, Harvard, IBM, Poughkeepsie

The adequacy of this characterization for science has essentially been established by work in logic and logical model theory, and we will not go into this here.

The notions of proper name and predicator are also clearly visible in recent linguistic theory. Thus it is said that strings such as "The dog is male" and "male dog" are transformationally related because both predicate that the individual referred to is both a "male" and a "dog." (The above is a deliberate over-simplification of actual linguistic statements).

Some computational linguists approach the problems of semantics by attempting to identify a finite set of elemental predicates (semantic primitives) which underlie meaning. In linguistic theory, the Fodor-Katz notion of semantic markers is a lucid embodiment of the search for semantic primitives. [2] Specific sets of such markers have been experimented with in some systems, for example by S. Cecatto.[3] Currently, Roger Schank at Stanford is developing these notions in a more detailed manner.[4] The question of whether such semantic primitives exist is moot or at least premature for discussion. Bolinger's brilliant article clearly indicated some perplexing problems relative to such semantic primitives. [5] Setting aside the theoretical aspects, our own views emphasize that, for practical systems operating at a relevant level of conceptual sophistication, the reduction of the high level notions involved to adequate perceptual and cognitive primitives is not only a tremendously complex task, but one not feasible from the point of view of practical computational efficiency and the relevant tasks at hand. The justification of this view is presented with examples later on in the paper.

Thus, the main aspect of the task at hand can be characterized in the following way. We wish to develop a system for communicating with the computer in, say, natural English addressed at selected universes of discourse. Each of these universes of discourse involves a set of entities and certain properties and relationships among them. Because of the diversity of applications of the system itself, we want our system to be independent of any particular subject matter but demand that it be readily adaptable to a wide variety of potential applications. A user must be able to select at will those properties and predicates that arise naturally from any new subject matter. The user's selection will be, of course, directly related to the collection of data or the cognitive model to which the system is to be applied. Communication, after all, rests upon confrontation of the communicants involved, be they human-to-human, or human-to-computer.

We have stressed earlier that, from a practical point of view, the properties and predicates germane to a given computational application of a language domain can hardly be reduced to a fixed set of universal primitives. A stronger point needs to be made. In this regard, it is the cogent selection of the high level, abstract properties and relationships that is at the core of the creative process.[6] Thus if a computer system is indeed to be "natural" to the user -- an intellectually sophisticated user -- such a system must accept new properties and predicates that subcategorize and interrelate his data in novel ways which are responsive to the user's own growing perception of his subject matter. In other words, the system must make room for conceptual creativity, since only in this way can it fill the user's need for "natural" language use.

From Sentences to Data: the Problem of Verbs

Of the two principal components of sentences, the noun phrases and the verb phrases, the former can be related to individuals, classes and relations within the data base in a straightforward way. Thus the meaning of "Boston", so far as the computer is concerned, is established by a link to that record in the data base where the data about Boston is located. The noun phrase: "mayor of Boston" is only slightly more complicated.

The problem of verbs and verb phrases is a more complex matter. As is well known, the verb (or verb phrase) can be considered central to the English sentence (and those of many other languages); it plays the central role in tying the various noun phrases of the sentence into a coherent whole. Respective to the data base, the problem is how to relate the verb to the individual/predicate structure of the records in the data base. The importance of this task and the desirability of a widely applicable solution is the central concern of this paper.

The semantic substrata on which computational schemes for handling verbs must currently be based are those of entities or individuals, on the one hand, and properties and predicates concerning these individuals, on the other. Although a verb may indeed carry more subtle aspects of meaning than can be encompassed by class membership and relations with other individuals, in all practicality it is with just such primitives that computational linguistic solutions addressed at data bases must currently work.

What does a verb do in a sentence? It performs a complex and temporally oriented predication of relationships existing among the entities identified by the noun phrases associated with it. [7]

For example, the verb "arrive" is associated with the relationship of location. It predicates that the subject of the clause is located in the place identified by whatever locative phrase is there. Further, that this relationship, namely one of location, was initiated at the time stipulated by the tense of the verb and the adverbs of time. Thus:

John arrived in Boston in June 1960.

can be paraphrased as:

Boston became the location of John in June 1960.

or alternately:

Boston was the location of John starting in June 1960.

The verb "come" in the sense of "arrive" can be handled in a similar manner, and the verbs "leave", "depart" in the opposite way, i. e. the ending of the "location" relationship is predicated. The verb "move" combines the meanings of (ending) and (beginning), and thus:

John moved from Boston to New York on July 27, 1971.

is paraphrased in terms of the associated predicates as:

Boston was John's location until July 27, 1971, and New York was John's location after July 27, 1971.

An excellent example of how the meaning of a verb can be paraphrased in terms of properties and predicates was given by Woods:

> "As an example, consider the semantic rule:
> (S: WRITE
>   (S: NP(MEM 1 PERSON))
>   (S: V-OBJ(AND(MEM 2 DOCUMENT) (EQU 1 WRITE)))
>   -- (PRED (AUTHOR: (#2 2) (#1 1))))
> This rule says that if the sentence has a subject which is a person, a verb "write", and an object which is a document, then the meaning of the sentence is computed by substituting the interpretations of the node numbered 1 in the first component (#1 1) and the node numbered 2 in the second component (#2 2) into the indicated places in the schema (AUTHOR(#2 2)(#1 1)) and treating it as a predicate (PRED). (S: WRITE is the name of the rule".) [8]

This treatment of the verb "write" appears to be the same as ours, i.e. the resulting paraphrase is: "author(Scott, Waverly)."

Fillmore also has analyzed verbs in terms of predicates, for instance in "Subjects, Speakers, and Roles": [9]

> "Certain verbs and adjectives seem to require inherently a given number of NPs in the sentences in which they take part. Another way of saying this is that certain verbs and adjectives seem quite naturally to be reconstructible as n-place predicates in formulations within the predicate calculus."

Thus, the problem of defining verbs seems reducable to relating the verb and noun phrases in a particular clause to the verb - associated relationship and the noun phrases. This kind of representation would be easier to achieve if the noun phrases in a clause were always in a given, statable order relative to the verb, and if the verbs were always transitive, that is, in the frame:

subject - verb - object.

But, to consider a very simple example, there are these various tranforms of:

"John owns the book":

The book is owned by John

Is the book owned by John

Does John own the book

The book John owns

John who owns the book

The notions of deep case grammar, developed by Fillmore, have been found very useful in handling these problems, not only syntactically, but primarily semantically, which is our concern here. [10]

In our application of Fillmore's theory the result of the syntactic analysis of a sentence is the identification of the verb phrase (including its tense modification), and, for each principal noun phrase in the sentence, an identification of a case relationship between that noun and the central verb. This is illustrated by the following sentences:

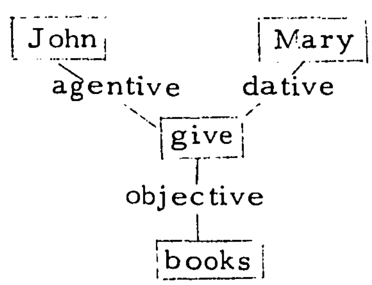
John gave books to Mary

John gave Mary books

Mary was given books by John

Books were given to Mary by John

all of which are surface formations for the following simple deep case analysis:



The deep case structure of the sentence, an appropriate definition of the verb, and data base referents for the various noun phrases represent the meaning of a given sentence.

Verbs in REL English

REL English grammar is currently based upon a deep case grammar. It recognizes the following seven cases:

agentive (AG)

objective (OJ)

instrumental (AI)

dative (DA)

locative (LO)

genitive (GE)

adverb of time (AT)

The result of parsing a sentence is its deep case structure in the above sense [11]

Our discussion of verb semantics hinges at this point on an explanation of a very important feature of the REL English which we have so far ignored. And that is that names, properties and predicates can be introduced into the language in a simple and direct way. Why this is important may not be immediately obvious. However, if we recall that verbs are indeed defined in terms of the underlying data entities - individuals, classes, and relationships among them - the significance is brought out. In general language use, it is not at all clear how verbs are understood. Theories so far advanced (cf. Chomsky and Fillmore) stress the relationships with nouns, in more or less abstract ways. In actual dealing with data bases, it becomes obvious that nouns are the stuff of verbs, i. e. verbs need to be related to the references actualized by nouns as either individuals, classes or relations among those. The traditional notion of verbs as indicating 'states' or 'events' is not to be taken lightly. The implication for their semantic reality derives from there. 'States' and 'events' are empty notions unless there are nouns of which the verbs 'state' or which they 'event'. Thus, also in terms of a data base, all the verbs do is to connect with nouns. They express processes on nouns. Therefore, if we are to use verbs, we need to use nouns which they will use. If we are to introduce verbs into language, we first need to introduce nouns. In REL English, they can be introduced in the following way:

John: =name

Ivenhoe: =name

owner: =relation

boy: =class

age: =number relation

These open the way for such statements as:

John is a boy.

John's age is 12.

John is the owner of Ivenhoe.

When it comes to verbs, and in order to be able to introduce them by definition, a mechanism is provided for declaring the case relationships that are to exist between the verb and the nouns. It consists of variables to be used as place holders, one for each case (e. g. , AG for agentive) in "paraphrasing" the meaning of the verb. Thus, if the relationship of "owner" is known to the system and the verb "own" is to be defined, the meaning of the verb would be paraphrased as follows:

own: =verb (AG is the owner of OJ)

The AG and OJ variables establish which case associated nouns play what roles. And, if subsequently, one were to ask:

Is Ivenhoe owned by John?

"Ivenhoe" would be identified as OJ and "John" would be identified as AG according to the deep case grammar rules. Therefore, the application of the above definition of "own" would yield (in an internalized form):

John is owner of Ivenhoe?

It will be noticed that the verb "own" was defined above in terms of the predicate "owner" which had previously been defined as a relation.

To illustrate a more complex situation, consider the meaning of the verb "give," as in "John gave Mary flowers."
It could be defined in the following way:

give: =verb (OJ will have been owned by AG before AT and
OJ will have been owned by DA after AT)

As illustrated in this definition, once a verb is defined, it may subsequently be used in the definition of new verbs; here the verb "own," used in defining "give".

Let us consider now the verb "sell," as in

"John sold the house to Mary in 1950 for $50,000"

We can define its meaning in the following way:

sell: =verb (AG gives OJ to DA at AT and the worth of AG is increased by AI and the worth of DA is decreased by AI)

assuming that "give" was defined as above. Is this, indeed, what is meant by the verb "sell"? Certainly there are aspects of the notion of 'selling' that are not entirely grasped by this definition such as that the change of "worth" of the parties involved is directly related to the exchange of ownership. However, let's make certain presuppositions concerning the data base underlying this definition and the subsequent questions in which the word "sell" may be used. Let us assume that this data base concerns a population of individuals and a set of items, further, that at any point in time it establishes who owns each item; finally that it gives the net worth of each of the individuals. In this case, the above definition fits the data base well; it is quite a natural definition when recording and quering data concerning transactions. A reverse analysis would, of course, hold for the verb "buy."

Let us consider now somewhat more extensively a specific contextual environment and related verb meaning definitions. This example aims at bringing out the preponderance of specific contextual settings, or the context-sensitivity of "natural" language.

A high school near Caltech has an enrollment of about 200 students and offers about 70 courses each semester. As a test case for an application of the REL System, we have attempted to see whether the REL System could be useful in scheduling courses. The data available consists of two files:

(a) for each student, a list of courses he or she may wish to take;

(b) categories of courses and specialties of teachers.

Thus, typical data input sentences are:

John Jones is a student.

English 2 is a course of John Jones.

Biology is a science course.

Helen Trent is a language teacher.

In scheduling courses, one would like to check on potential conflicts. To this end one might ask:

What students take biology and geometry?

The problem is to define "take" so that it may be used in this, and similar queries, with the appropriate meaning. A solution is to define "take" as follows:

take: =verb (OJ is a course of AG)

Given the above, one can ask:

What students do not take any science course?

What courses are taken by less than three students that take history 2?

As conflicts are resolved the notion of the period of the day may be defined, for instance:

third period course: =class

and courses assigned to it; e. g.:

Math 2 is a third period course.

And if one wants to ask, for instance:

What science teacher does not give any third period course?

one would define "give" as:

give: =verb (AG is the teacher of OJ)

It will be noticed here that this definition of "give" differs radically from the definition previously discussed concerning questions of ownership and transactions. The latter fits the context of the school curriculum whereas the former fits quite different contexts.

Conclusions

The essential point we wished to stress in this paper is that the meaning of a sentence depends on the contents of the data base to which it refers.

Linguistically, the tying string of a sentence is its verb - that elusive VP which ties the distinct NPs into a coherent pattern. And this sentence pattern takes on meaning as it fits within the broader fabric supplied by context and reality. It is doubtful that there is any single method for analyzing the complex webs that verbs set up among their associated nouns in any sentence. But the examples of the verbs "give" and "take" as discussed here may perhaps shed some light on the importance and inevitability of context in language use. These examples show how these same verbs radically shift their semantic content with shifts in the contexts in which they are used.

References and Notes

[1] Bross, I. D. J., P. A. Shapiro, and B. B. Anderson, "How Information Is Carried in Scientific Sub-Languages," Science, June 23, 1972, vol. 176, p. 1306.

[2] Fodor, J. A. and J. J. Katz, "The Structure of a Semantic Theory," The Structure of Language, Prentice-Hall, Englewood Cliffs, N. J., 1964, 479-518.

[3] Ceccato, S., "Towards the Automation of Mental Activities," 3rd International Congress on Cybernetics, Namur, 1961, 473-476.

[4] Schank, R., "The Role of Theoretical Linguistics in Text Processing," ONR Conference on Text Processing and Scientific Research, Pasadena, November 1972.

[5] Bolinger, D., "The Atomization of Meaning," Language 41, No. 4 (1965), 555-573.

[6] The notion of the creative process in language has received much attention and emphasis, primarily on the part of N. Chomsky, with reference to sentence structure. We address ourselves to the more general, and particularly ill-understood, phenomenon of language use for communication. The creative process of concept formation is of utmost importance here.

[7] The discussion that follows involves temporal relations, which are among the most difficult aspects of verb phrases. The intricate manner of handling temporal clauses cannot be gone into in this paper. The presentation here is a somewhat simplified picture of the way we currently handle reference to time. A publication on the topic is forthcoming.

[8] Woods, W. A., R. M. Kaplan and B. Nash-Webber, The Lunar Sciences Natural Language Information System, Bolt Beranek and Newman, Cambridge, June 1972, p. 220.

[9] Fillmore, C., "Subjects, Speakers and Roles," Ohio State University Working Papers in Linguistics, No. 4, Columbus, 1970, p. 39.

[10] Fillmore, C., "The Case for Case," Bach, E. (ed.), Universals in Linguistic Theory, Holt, Rinehart and Winston, New York, 1968, 1-90.

[11] An example of a sentence involving all cases would be this: John (AG) bought flowers (OJ) for Mary (DA) in Boston (LO) from Pete (GE) with stolen money (AI) in 1967 (AT). The theory underlying our cases is still not clear to us, but they seem necessary to handle sentences such as above. The (GE) may be particularly controversial and probably might be better expressed in terms of the "Source" case.