COGNITIVE SCIENCE 4, 209–241 (1980)

What Does It Mean to Understand Language?

TERRY WINOGRAD
Stanford University

INTRODUCTION

In its earliest drafts, this paper was a structured argument, presenting a comprehensive view of cognitive science, criticizing prevailing approaches to the study of language and thought and advocating a new way of looking at things. Although I strongly believed in the approach it outlined, somehow it didn't have the convincingness on paper that it had in my own reflection. After some discouraging attempts at reorganization and rewriting, I realized that there was a mismatch between the nature of what I wanted to say and the form in which I was trying to communicate.

The understanding on which it was based does not have the form of a carefully structured framework into which all of cognitive science can be placed. It is more an orientation—a way of approaching the phenomena—that has grown out of many different experiences and influences and that bears the marks of its history. I found myself wanting to describe a path rather than justify its destination, finding that in the flow, the ideas came across more clearly. Since this collection was envisioned as a panorama of contrasting individual views, I have taken the liberty of making this chapter explicitly personal and describing the evolution of my own understanding.

My interests have centered around natural language. I have been engaged in the design of computer programs that in some sense could be said to "understand language," and this has led to looking at many aspects of the problems, including theories of meaning, representation formalisms, and the design and construction of complex computer systems. There has been a continuous evolution in my understanding of just what it means to say that a person or computer "understands," and this story[1] can be read as recounting that evolution. It is

[1]This is a "story" because like all histories it is made up. In an attempt to make sense of the chaos of past events one imposes more of a sense of orderliness than they deserve. Things didn't actually happen exactly in this order, and the events contain inconsistencies, throwbacks, and other misfortunes that would make it much harder to tell.

long, because it is still too early to look back and say "What I was *really* getting at for all those years was the one basic idea that . . ." I am too close and too involved in its continuation to see beyond the twists and turns. The last sections of the paper describe a viewpoint that differs in significant ways from most current approaches, and that offers new possibilities for a deeper understanding of language and a grasp on some previously intractable or unrecognized problems. I hope that it will give some sense of where the path is headed.

2. UP THROUGH SHRDLU

The Background

In the mid 1960s, natural language research with computers proceeded in the wake of widespread disillusionment caused by the failure of the highly touted and heavily funded machine translation projects. There was a feeling that researchers had failed to make good on their early confident claims, and that computers might not be able to deal with the complexities of human language at all. In AI research laboratories there were attempts to develop a new approach, going beyond the syntactic word-shuffling that dominated machine translation and other approaches based on key word search or statistical analysis. It was clear that for effective machine processing of language—whether for translation, question answering, or sophisticated information retrieval—an analysis of the syntactic structures and identification of the lexical items was not sufficient. Programs had to deal somehow with what the words and sentences meant.

There were a number of programs in this new vein described in the early collections of AI papers.[2] Each program worked in some very limited domain (baseball scores, family trees, algebra word problems, etc.) within which it was possible to set up a formal representational structure corresponding to the underlying meaning of sentences. This structure could be used in a systematic reasoning process as part of the overall language comprehension system. The model of language understanding that was implicit in those programs and in many AI programs since then is illustrated in Figure 1.

This model rests on some basic assumptions about language and representation:

1. Sentences in a natural language correspond to facts about the world.
2. It is possible to create a formal representation system such that
   (a) For any relevant fact about the world there is a corresponding structure in the representation system;

[2]Green et al. and Lindsay in Feigenbaum and Feldman (1963); Black, Bobrow, Quillian, and Raphael in Minsky (1967).

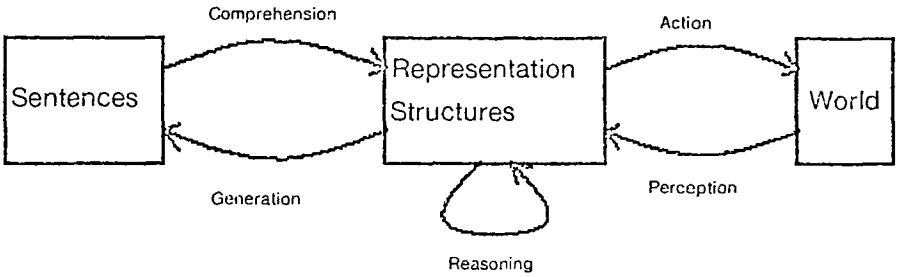

Figure 1. Basic AI model of language understanding.

(b) There is a systematic way of correlating sentences in natural language with the structure in the representation system that correspond to the same facts about the world; and
(c) Systematic formal operations can be specified that operate on the representation structures to do "reasoning." Given structures corresponding to facts about the world, these operations will generate structures corresponding to other facts, without introducing falsehoods.

This somewhat simplistic formulation needs some elaboration to be comprehensive. It is clear, for example, that a question or command does not "correspond to facts" in the same manner as a statement, and that it is unlikely that any actual reasoning system will be error-free. We will discuss some of these elaborations later, but for a first understanding, they do not play a major role.

The critical element in this model that distinguishes it from the pre-AI programs for language is the explicit manipulation of a formal representation. Operations carried out on the representation structures are justified not by facts about language, but by the correspondence between the representation and the world being described. This is the sense in which such programs were said to "understand" the words and sentences they dealt with where the earlier machine translation programs had "manipulated them without understanding."

This general model was not a new idea. It corresponds quite closely to the model of language and meaning developed by philosophers of language like Frege, drawing on ideas back to Aristotle and beyond. There has been a good deal of flag waving at times about the ways in which the "artificial intelligence paradigm" is new and superior to the older philosophical ideas. In large degree (some exceptions are discussed later) this has been a rather empty claim. As Fodor (1978) has pointed out, to the extent they are clearly defined, AI models are generally equivalent to older philosophical ones. A formal logical system can play the role of a representation system as described in the figure, without being explicit about the nature of the processing activity by which reasoning is done.

In fact, AI programs dealing with language do not really fit the model of Figure 1, since they have no modes of perception or action in a real world.

Although they converse about families, baseball or whatever, their interaction is based only on the sentences they interpret and generate. A more accurate model for the programs (as opposed to the human language comprehension they attempt to model) would show that all connection to the world is mediated through the programmer who builds the representation. The reason that "dog" refers to dog (as opposed to referring to eggplant parmesan or being a "meaningless symbol") lies in the intention of the person who put it in the program, who in turn has knowledge of dogs and of the way that the symbols he or she writes will be used by the interpreter. This difference is important in dealing with questions of "background" discussed later.

SHRDLU

SHRDLU (Winograd, 1972) was a computer program for natural language conversation that I developed at MIT between 1968 and 1970.[3] The program carried on a dialog (via teletype) with a person concerning the activity of a simulated "robot" arm in a tabletop world of toy objects. It could answer questions, carry out commands, and incorporate new facts about its world. It displayed the simulated world on a CRT screen, showing the activities it carried out as it moved the objects around.

SHRDLU had a large impact, both inside and outside the field, and ten years later it is still one of the most frequently mentioned AI programs, especially in introductory texts and in the popular media. There are several reasons why so many people (including critics of AI, such as Lighthill (1973)) found the program appealing. One major factor was its comprehensiveness. In writing the program I attempted to deal seriously with all of the aspects of language comprehension illustrated in the model. Earlier programs had focussed on one or another aspect, ignoring or shortcutting others. Programs that analyzed complex syntax did not attempt reasoning. Programs that could do logical deduction used simple patterns for analyzing natural language inputs. SHRDLU combined a sophisticated syntax analysis with a fairly general deductive system, operating in a "world" with visible analogs of perception and action. It provided a framework in which to study the interactions between different aspects of language and emphasized the relevance of nonlinguistic knowledge to the understanding process.

Another factor was its relatively natural use of language. The fact that person and machine were engaged in a visible activity in a (pseudo-)physical world gave the dialog a kind of vitality that was absent in the question-answer or problem-solution interactions of earlier systems. Further naturalness came from the substantial body of programs dealing with linguistic phenomena of conversation and context, such as pronouns ("it," "that," "then," etc.), substitute

[3]Winograd (1971) was the original dissertation. Winograd (1972) is a rewritten version that owes much to the editing and encouragement of Walter Reitman. Winograd (1973) is a shortened account, which also appears in various reworkings in several later publications.

nouns ("a green *one*"), and ellipsis (e.g., answering the one-word question "Why?"). Dialog can be carried on without these devices, but it is stilted. SHRDLU incorporated mechanisms to deal with these phenomena in enough cases (both in comprehension and generation) to make the sample dialogs feel different from the stereotype of mechanical computer conversations.

In the technical dimension, it incorporated a number of ideas. Among them were:

1. Use of a reasoning formalism (MicroPlanner) based on the "procedural embedding of knowledge." Specific facts about the world were encoded directly as procedures that operate on the representation structures, instead of as structures to be used by a more general deductive process. The idea of "procedural embedding of knowledge" grew out of early AI work and had been promoted by Hewitt (1971). SHRDLU was the first implementation and use of his Planner language. The difference between "procedural" and "declarative" knowledge has subsequently been the source of much debate (and confusion) in AI.[4] Although procedural embedding in its simplistic form has many disadvantages, more sophisticated versions appear in most current representation systems.
2. An emphasis on how language triggers action. The meaning of a sentence was represented not as a fact about the world, but as a command for the program to do something. A question was a command to generate an answer, and even a statement like "I own the biggest red pyramid" was represented as a program for adding information to a data base. This view that meaning is based on "imperative" rather than "declarative" force is related to some of the speech act theories discussed below.
3. A representation of lexical meaning (the meaning of individual words and idioms) based on procedures that operate in the building of representation structures. This contrasted with earlier approaches in which the lexical items simply provided (through a dictionary lookup) chunks to be incorporated into the representation structures by a general "semantic analysis" program. This was one of the things that made it possible to deal with conversational phenomena such as pronominalization. Some equivalent device is present in most current natural language programs, and there is a formal analog in the generality of function application in Montague Grammar formalizations of word meaning.
4. An explicit representation of the cognitive context. In order to decide what a phrase like "the red block" refers to, it is not sufficient to consider facts about the world being described. There may be several red blocks, one of which is more in focus than the others because of having been mentioned or acted on recently. In order to translate this phrase into the appropriate representation structure, reasoning must be done using representation structures corresponding to facts about the text preceding the phrase, and structures corresponding to facts about which objects are "in focus."

The attempt to deal with conversational phenomena called for an extension to the model of language understanding, as illustrated in Figure 2. It includes additional structures (as part of the overall representation in the language understander) labelled "model of the text" and "model of the speaker/hearer." The label

[4]See Winograd, (1975) for discussion.

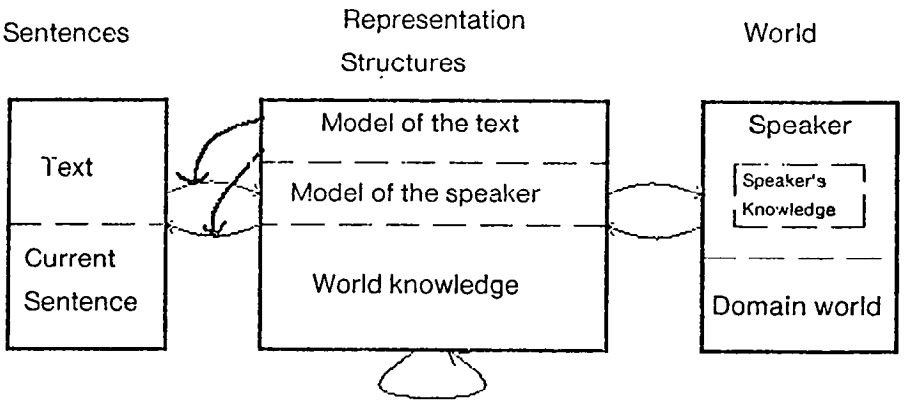

Figure 2. Extended AI model of language understanding.

"model of the speaker" was chosen to reflect the particular approach taken to the problem. It is assumed that inferences about which objects are in focus (and other related properties) can be made on the basis of facts about the knowledge and current internal state (presumably corresponding to representation structures) of the other participant in the conversation. The question "could I use this phrase to refer to object X?" is treated as equivalent to "if I used this phrase would the hearer be able to identify it as naming object X?" On the other side, "what does he mean by this phrase?" is treated as "what object in his mind would he be most likely to choose the phrase for?"

In addition to reasoning about the domain world (the world of toy blocks), the system reasons about the structure of the conversation and about the hypothesized internal structure and state of the other participant. In SHRDLU, this aspect of reasoning was not done using the same representation formalism as for the domain world, but in an *ad hoc* style within the programs. Nevertheless, in essence it was no different from any other reasoning process carried out on representation structures.[5]

3. SEEING SOME SHORTCOMINGS

SHRDLU demonstrated that for a carefully constrained dialog in a limited domain it was possible to deal with meaning in a fairly comprehensive way, and to achieve apparently natural communication. However, there were some obvious problems with the approach, summarized here and discussed below:

[5]For a more elaborate version of this model, along with many examples of conversational phenomena not handled by SHRDLU or any other existing computer system, see Winograd (1977a). Winograd (in preparation) presents an overview of syntactic and semantic structures within a viewpoint drawn from this model.

1. The explicit representation of speaker/hearer internal structure was *ad hoc*, and there was no principled way to evaluate extensions.
2. The notion of word definition by program, even though it opened up possibilities beyond more traditional logical forms of definition, was still inadequate.
3. It took rather strained reasoning to maintain that the meaning of every utterance could be structured as a command to carry out some procedure.
4. The representation and reasoning operations seemed inadequate for dealing with common-sense knowledge and thought reflected in language.

The Internal Structure

In building a simpler system as illustrated in Figure 1, the programmer is creating a model of the language comprehension process. In creating the representation structures corresponding to facts about the domain, he or she is guided by an idea of what is true in the domain world—in representing facts about blocks, one can draw on common-sense knowledge about physical objects. On the other hand, in trying to create structures constituting the model of the speaker/hearer as in Figure 2, there is no such practical guide. In essence, this model is a psychological theory, purporting to describe structures that exist in the mind. This model is then used in a reasoning process, as part of a program whose overall structure itself can be thought of as a hypothesis about the psychological structure of a language understander.

Experimental psychology provides some suggestive concepts, but little else of direct use. A language comprehension system depends on models of memory, attention, and inference, all dealing with meaningful material, not the well-controlled stimuli of the typical laboratory experiment. Research in cognitive psychology has focussed on tasks that do not clearly generalize to these more complex activities. In fact, much current psychological research on how people deal with meaningful material has been guided by AI concepts rather than the other way around.

The problem is hard to delimit, since it touches on broad issues of understanding. In SHRDLU, for example, the program for determining the referent of a definite noun phrase such as "the block" made use of a list of previously mentioned objects. The most recently mentioned thing fitting the description was assumed to be the referent. Although this approach covers a large number of cases, and there are extensions in the same spirit which cover even more, there is a more general phenomenon that must be dealt with. Winograd (1974a) discusses the text "Tommy had just been given a new set of blocks. He was opening the box when he saw Jimmy coming in."

> There is no mention of what is in the box—no clue as to what box it is at all. But a person reading the text makes the immediate assumption that it is the box which contains the set of blocks. We can do this because we know that new items often come in boxes, and that opening the box is a usual thing to do. Most important, we assume that we are receiving a connected message. There is no reason why the box has to be

> connected with the blocks, but if it weren't, it couldn't be mentioned without further introduction. (Winograd, 1974a)

Important differences in meaning can hinge on subtle aspects of the speaker/hearer model. For example, in the first sentence below, it is appropriate to assume that the refrigerator has only one door, while in the second it can be concluded that it has more than one. On the other hand, we cannot conclude from the third sentence that the house has only one door.

When our new refrigerator arrived, the door was broken.
When our new refrigerator arrived, a door was broken.
When we got home from our vacation, we discovered that the door had been broken open.

The problem, then, is to model the ways in which these connections are made. In general this has led to an introspective/pragmatic approach. Things get added to the representation of the speaker/hearer because the programmer feels they will be relevant. They are kept because with them the system is perceived as performing better in some way than it does without them. There have been some interesting ideas for what should be included in the model of the speaker/hearer, and how some of it might be organized[6] but the overall feeling is of undirected and untested speculation, rather than of persuasive evidence or of convergence towards a model that would give a satisfactory account of a broad range of language phenomena.

Word Definition

The difficulty of formulating appropriate word definitions was apparent even in the simple vocabulary of the blocks world and becomes more serious as the domain expands. In SHRDLU, for example, the word "big" was translated into a representation structure corresponding to "having X,Y, and Z coordinates summing to more than 600 units (in the dimensions used for display on the screen)." This was clearly an *ad hoc* stopgap, which avoided dealing with the fact that the meaning of words like "big" is always relative to an expected set. The statement "They were expecting a big crowd" could refer to twenty or twenty thousand, depending on the context. By having word definitions as programs, it was theoretically possible to take an arbitrary number of contextual factors into account, and this constituted a major departure from more standard "compositional" semantics in which the meaning of any unit can depend only on the independent meanings of its parts. However, the mere possibility did not provide a guide for just what it meant to consider context, and what kind of formal structures were needed.

On looking more closely, it became apparent that this problem was not a special issue for comparative adjectives like "big," but was a fundamental part

[6]See for example Schank and Abelson (1975), Hobbs (1978). and Grosz (1980).

of the meaning of most words. Linguists have pointed out that a natural language categorization cannot be equated with a finite set of logical predicates to be applied to the situation.[7] The applicability of a word depends on an understanding of the purposes of speaker and hearer. Winograd (1976) discusses the meaning of "bachelor." In classical discussions of semantics, "bachelor" has been used as an example of a word with a clear paraphrase in more elementary terms—"unmarried adult human male."[8] But if someone refers to a person as a "bachelor" in normal conversation, much more is meant. It is inaccurate if used in describing the Pope or a member of a monogamous homosexual couple, and might well be used in describing an independent career woman.

The issue is not that the definition of bachelor is complex and involves more terms than usually accounted for. There is no coherent checklist of any length such that objects meeting all of its conditions will consistently be called "bachelors" and those failing one or more of them will not. The question "Is X a bachelor?" cannot be answered without asking "Why do you want to know?". Of course, it is possible to create artificial strict definitions, but these do not account for the normal use of language. When we move to a larger vocabulary, the problem becomes even more obvious. Each of the nouns in the sentence "The administration's dishonesty provoked a crisis of confidence in government" raises a significant problem of definition, and it is clear that purpose and context play a major role in determining what will be called a "crisis," "dishonesty," or even in deciding just what constitutes an "administration."

The problem of considering the contextual background in the meaning of words is not solved simply by moving from a "declarative" form of definition to a "procedural" one, as in SHRDLU. The activity of a program in any bounded amount of time can be described in purely logical terms, and a logical formula can be written which can be proved true in exactly those cases where the program would return the answer "yes" and false when it would say "no."

Meaning as Command

SHRDLU was based on a formalism in which the meaning of a sentence was represented as a command to carry out some action. A question is a command to generate a sentence satisfying a set of constraints, and a statement is a command to add a formula to the data base. This shift of viewpoint from meaning-as-statement to meaning-as-command provided some interesting ways of talking about sentences, but in its naive form it is clearly unworkable.

The good part of the idea was the view of an utterance as triggering some kind of activity in the hearer. The bad part of the idea was the analogy with

[7]See, for example, Fillmore (1975) and Labov (1973).

[8]This is, of course, only one of its definitions. Others, as pointed out by Katz and Fodor (1964), relate to fur seals and chivalry.

computer programming languages, in which there is a direct correspondence between the linguistic form and the sequence of activities to be carried out. In the case of natural language, much more is going on. Both speaker and hearer are engaged in ongoing processes of trying to make sense of their conversation and the world they inhabit. The interpretation of utterances is only one of many activities, and interacts with perception, reasoning, memory, and all the other aspects of cognition mentioned by Norman (this volume). When I utter a sentence, I have no way of anticipating in any detail the processing it will invoke in the hearer. It clearly includes much more than simply obeying or storing away a fact. The following simple examples illustrate some of what goes on.

1. Tom has never failed a student in Linguistics 265.
2. I'm sorry I missed the meeting yesterday. My car had a flat tire.
3. There's an animal over there in the bushes.

Sentence 1 is true in many circumstances, including the one in which Tom has never taught Linguistics 265. However, in ordinary conversation, the hearer makes the additional implication that Tom has taught the course, and is justified in accusing the speaker of bad faith if the implication is not warranted. Similarly in sentence 2, the hearer assumes that there is a coherence to the events being described. If the second sentence were "There are fifteen million people in Mexico City" the hearer would be puzzled, and if the flat tire had nothing to do with missing the meeting (even though it actually did happen), the speaker is practicing deception.

Sentence 3 is a more subtle case. If the hearer looks over and sees a dog in the bushes, and finds out that the speaker knew it was a dog, he or she will feel that the statement was inappropriate, and might say "If you knew it was a dog, why didn't you say so?". On the other hand, the statement "There's a dog over there in the bushes" is perfectly appropriate even if both speaker and hearer know that it is a beagle,[9] and sentence 3 would be fine for a dog if it were a response to something like "There are no animals anywhere around here."

The common element in all of these is that the "meaning" for a hearer is the result of a complex process of trying to understand what the speaker is saying and why. In effect, every statement is the answer to a question, which may be implicit in the context. Its meaning depends as much on the question as on the form of the answer. There is an important germ of truth in saying that the meaning of a sentence "is" the process it invokes, but this view is not compatible with a formal compositional semantics of the kind that has generally interested linguists and philosophers. What is needed is an understanding of meaning-as-triggering, which deals with the interaction between the utterance and the full range of cognitive processes going on in the language user.

[9]These phenomena are related to the properties of human categorization systems studied by Rosch (1975).

[10]See the discussion in Fodor (1978).

Natural Reasoning

In looking at any significant sample of natural language, it becomes quickly apparent that only a small fraction of human "reasoning" fits the mold of deductive logic. One is often presented with a fragmentary description of some object or situation, and on the basis of knowledge about what is "typical" jumps to a number of conclusions that are not justifiable as logical deductions, and may at times be false. Most AI systems have been based (either explicitly or unwittingly) on a notion of deduction that does not account for this style of reasoning. In the "opening the box" example quoted above, it was assumed that "the box" was the one in which the blocks arrived even though this is not rigorously deducible from the text, or even from knowledge of the world. This kind of inference is a predominant aspect of reasoning and one which calls for formal systems having very different properties from deductive logic.[11]

4. KRL: TOWARD A NEW FORMALISM

My recognition of these problems with the approach underlying SHRDLU came at the time (1972–73) I participated in a series of discussions in the Artificial Intelligence Lab at MIT about the problems of natural reasoning. These discussions led among other things to Minksy's (1975) formulation of "frames" and to my own attempts to make these intuitions more precise (1974b, 1975). The next major step toward clarifying them was in the specification and construction of KRL (Knowledge Representation Language), a project done jointly with Daniel Bobrow at Xerox PARC.[12] The project has included two rounds of design and implementation, with a limited amount of testing in small applications. Each implementation has covered only a fraction of the overall intended features of the language, and the discussion below (except where otherwise indicated) deals with the basic concepts, not the current implentations.

The Formalism

Viewed as a language for use in building systems of the kind described in Figures 1 and 2, KRL was intended to have the following properties:

1. *Resource-limited processing as a basis for reasoning.* In any act of interpretation or reasoning, a system (biological or computer) has a finite quantity of processing resources to expend. The nature of these resources will be affected by the details of the processor, its environment, and previous history. The outcome of the process is determined by the interaction between the structure of the task and the allocation of process-

[11] Winograd (1980) and several other papers in the same issue of *Artificial Intelligence* deal with problems of "nonmonotonic logic" related to this problem.

[12] See Bobrow and Winograd (1977).

ing. The ability to deal with partial or imprecise information comes from the ability to do a finite amount of processing, then jump to a conclusion on the basis of what has happened so far.

2. *Separation of logical form from memory form.* In order to build a theory of resource use in processing, it is necessary to consider the details of how memory processes operate. Two pieces of information may be logically equivalent but differ in how easily they can be accessed or stored. Most representation systems have either ignored memory properties (as in most theories of formal logic), or assumed that they followed from the logical properties of the information. In KRL, stored knowledge has separable dimensions of logical content and memory "chunking." The memory structure depends not just on the logical content, but also potentially the particular history of the way it was entered into the structures.
3. *The integration of meta-knowledge.* In order to use information about the knowledge structures in a systematic way, the system needs a capacity of partial self-description. Formal structures can be used to represent objects in an "outside world" such as blocks, other structures of the same form can be used to represent those structures in turn, and so on indefinitely. This makes it possible to formulate algorithms for resource-limited processing in the same language that is used for other representation.
4. *An overall control structure based on matching.* The basic organization of KRL provides a framework for an extended "matching" process in which the choice of what to do at each point is driven by an attempt to match stored "prototypes" against the current inputs. This matching can involve complex deductions.

The development of KRL led to deeper thinking about the nature of representation formalisms and the reasoning operations done with them. Looking from one point of view, a KRL system can be thought of as a purely logical formal system—it operates with precise rules on well-defined structures as does any computer program. At the same time, there is another view from which the system can be seen as carrying out a kind of informal reasoning—one that comes to conclusions based on partial evidence, makes assumptions about what is to be expected in typical cases, and leaves open the possibility of mistake and contradiction. A KRL system (or any system using resource-limited reasoning) can draw some conclusion, then reverse it on the basis of further reasoning.

The key to this paradox lies in the system's use of formal rules that make reference to the structure of the system of itself. In coming to a conclusion about some world being represented (such as a world of toy blocks,) a KRL system can come to conclusions on the basis not only of statements about the world, but also on the basis of the form of its structures (for example, concluding something is false because it is normally false, and its truth in this case does not emerge in some bounded amount of reasoning). There is a fundamental philosophical and mathematical difference between truth-based systems of logic, and process-based systems like KRL.[13] One of the major theoretical directions of the KRL effort is to make this difference clear and understand its consequences.

[13]These issues are discussed at length in Winograd (1980).

Reasoning and Language

KRL was designed to make possible a different formal basis for the definition of words in natural language. In a standard deductive system, the definition of a word is a formula specifying a set of necessary and sufficient conditions for its applicability. Some terms are designated as *primitive*, and the rest are defined in terms of them. Although nonprimitive terms can appear in definitions, there can ultimately be no *circular definitions*—every nonprimitive term can be expanded through its definition (and the definitions of the terms appearing in it, recursively) to a formula containing only connectives and primitives. In deciding whether a word applies to an object, this expansion is carried out and the primitive terms are checked against the case at hand. If (and only if) they all check out, the word applies.

Formalisms like KRL provide another way of looking at this.[14] A word is associated with a "prototype" in the representation formalism. This prototype (like a definition) includes a description of the object in other terms. However, unlike a definition, this further description is not taken to be sufficient or necessary for determining the applicability of the prototype. It can include things that are typical (but not always the case) or that are relevant only in some contexts. In deciding whether a word applies to an object, the reasoning system compares these further descriptions to what is known about the object. It does so in a resource-dependent way, possibly looking at only some of the description, and choosing which to do on the basis of context. After a limited amount of processing, it makes a decision (as to whether the word applies) on the basis of what has been examined so far.

A process of this type has the potential to treat word meanings in the open-ended way discussed for the "bachelor" example above. Although there is a "checklist" of further descriptions, the process of checking is context dependent and limited. It is possible to imagine strategies for deciding which items on the list to examine, depending on the current purposes and background. The KRL framework does not specify how this is to be done in detail, but through the combination of resource-limited reasoning and meta-description it provides tools that can be used.

Similarly, these tools can be applied to the other problems raised in the critique of SHRDLU above. The whole concept of resource-limited reasoning grew out of an analysis of the kinds of "natural reasoning" that cannot be done with a straightforward deductive system. The idea of "self-description" was partly an attempt to provide an explicit way of allowing the system to have a model of itself and, by reflection, of a hypothesized dialog partner. The matching framework provides a way to treat inputs to the system as elements of an overall

[14]The problem of word definition, and the notions of "prototype" and "primitive" are discussed at length in Winograd (1976, 1978) and Bobrow and Winograd (1979).

pattern being matched, rather than as items to be stored or commands to be carried out.

With all of these KRL provided only a framework. It was designed to facilitate exploration of these problems, not to embody solutions to them. In that sense it is a kind of "calculus" of natural reasoning, just as the predicate calculus provides a formal basis for systems of deduction and the differential calculus serves as language for describing physical systems. Having this calculus is not the same as having a solution to the specific problems, or a theory of language understanding; but it is a potentially valuable tool. The experience in using it so far is difficult to evaluate. No substantial project has been done that uses its capabilities to really address the kinds of problems that motivated its creation. In trying to use it more fully, technical difficulties (including everything from slow execution speed to user interfaces) are intertwined with conceptual difficulties (which are hard to isolate clearly). As discussed below, there are deep questions of just what could be expected if KRL were to "succeed," but in any event it is clear that it cannot realistically be evaluated in its current implementations.

5. THE HERMENEUTIC CONNECTION

At the same time that KRL was being developed I took part in a series of informal discussions in Berkeley about the nature of language and cognition. These discussions included philosophers, linguists, and computer scientists and ranged from the narrowest technical details to the broadest philosophical concerns. They raised questions about what it was that KRL (and all other computer representations) were claimed to achieve. Among the works discussed were those of Maturana (1977) on the biological foundations of cognition, and the literature on hermeneutics.[15] This paper is not the place for a detailed exposition of these ideas,[16] but there were some elements of both that were applicable to the problem of building programs that interacted in natural language. The following paragraphs attempt to lay them out in outline, but for a more thorough understanding, it is necessary to go to the original sources.

The Nervous System as a Closed, Plastic, Structure-determined System

Maturana proposes an understanding of the nervous system that is not built around the usual notions of input, output, memory, perception, etc. He adopts instead an orientation toward it as a system of components whose activities

[15]Gadamer (1976) was an important source, and Palmer (1969) is an excellent overall introduction.

[16]Flores and Winograd (in preparation) does so at some length.

trigger further activity within the system. The system is "structure determined" in that its activity at any moment is fully determined by the structure (or state) at that moment. It is "plastic" in that its structure can be changed by the activity, so that its structure at any moment is a product of the entire previous history of activity and changing structure. It is "closed," in the sense that the system can do only that which is determined by its own structure and activity—its action cannot be understood as a reflection of an external world it perceives.

At first reading, for most people this approach seems bizarre, apparently denying the obvious fact that we see, hear, and generally perceive a world outside of our own nervous system. It is not a denial, but a change of stance. Instead of looking at vision as a mapping of external reality onto mental structures, we can look at it biologically as a change to the structure of the nervous system, in particular to the chemical and electrical properties of various cells in the retina. The subjective introspection is that we "see something," but from a neurophysiological standpoint, there is a structure-determined causal network in which "perturbations" to the structure of the system lead to patterns of activity that are different from those that would have happened with different perturbations. The focus is shifted away from the structure of the phenomena that led to the perturbations toward the structure of changes in the ongoing activity of the system as it is perturbed.

This view meshed well with the "triggering" view of language understanding described above. An utterance is neither a description of something in the world, nor a command specifying what the hearer is to carry out. It is a "perturbation" to an active ongoing cognitive system that is trying to make sense of things. The central questions to be asked of an utterance are those dealing with the changes in activity that it triggers, not with its correspondence to a world it describes.

The Hermeneutic Circle

Hermeneutics is the study of interpretation. It began as a discipline with the problem of interpreting sacred texts, but has come to encompass not only the interpretation of language, but also the larger understanding of how we interpret the world in which we live. One of the fundamental insights of hermeneutics is the importance of *pre-understanding*. In any situation where we are interpreting language, we begin with a system of understanding that provides a basis within which to generate an interpretation. This pre-understanding in turn arises and evolves through the acts of interpretation. This circle, in which understanding is necessary for interpretation, which in turn creates understanding, is called the *hermeneutic circle*.

> But there is a contradiction here. How can a text be understood, when the condition for its understanding is already to have understood what it is about? The answer is that somehow, by a dialectical process, a partial understanding is used to understand still further, like using pieces of a puzzle to figure out what is missing . . . A certain

> pre-understanding of the subject is necessary or no communication will happen, yet that understanding must be altered in the act of understanding. (Palmer, 1969)

The parallel with the issues of reasoning with frame representations should be obvious. The set of stored schemas in a system is its "pre-understanding." The use of these schemas affects the interpretation of what is said:

> In comprehension, the set of stored schemas is actively used in a process of "pattern recognition." The hearer assumes that the discourse is made up of instances of known discourse and reasoning patterns. . . . Some feature of an utterance, together with the current context, can trigger a hypothesis that an instance of some particular schema is being conveyed. This hypothesis is tested by attempting to fit other parts of the utterance in as pieces of the hypothesized schema. As a result the way in which the input is analyzed can be controlled (or biased) by the fact that it is being processed as part of looking for an instance of a specific hypothesized schema. (Winograd, 1977a)

Although little has been said within AI and cognitive science about how new schemas are generated, it is clear that it must be the result of a history of previous interactions, each of which is mediated through other schemas in a hermeneutic circle.[17] In emphasizing the relationship between an act of interpretation and the history of previous interpretations by the system, hermeneutics raises some of the same questions as Maturana's approach to plastic, structure-determined systems.

6. FOUNDERING ON THE OPPOSITE SHOALS

My course to this point was a gradual steering away from the logical-deductive model and its view of language understanding based on objective truth. Starting with the essentially conventional semantic underpinnings of SHRDLU, I had become more and more concerned with the properties of language and thought that could not easily be embodied in traditional logical systems. Maturana and Gadamer provided some philosophical wind for the rudimentary KRL sails, and this new direction seemed ripe for exploration. Structure-dependent resource-limited reasoning offered a way of formalizing natural kinds of inference, and its dependence on the specific past history of the language understander opened up room for systematically dealing with the "nonlogical" phenomena of language.

In many ways, this direction is still open for development and there are many areas to be explored. But early *en route* I was exposed to some troubling questions about the rocks we faced on the other shore. Two basic issues stood out: the problem of subjective relativism and the problem of representation. In this section, we will lay out these problems without proposing solutions. The directions described in later sections were guided in part by an attempt to solve them.

[17]The development of schemas is of growing concern in AI, as indicated by Minsky (this volume) and Schank (this volume.)

Subjective Relativism

The first issue can be described in simplistic terms as a dispute about two different starting points for understanding language:

> *Objectivity:* an utterance has meaning by virtue of corresponding to a state of affairs. We approach the study of language by analyzing how the structures of utterances correspond systematically to the states of affairs they describe.
>
> *Subjectivity:* an utterance has meaning by virtue of triggering processes within a hearer whose cognitive structure depends on prior history and current processing activity. We approach the study of language by analyzing the nature of those cognitive structures and activities.

The examples given above suggest that a subject-dependent view must be taken even in such seemingly objective issues as the appropriate use of "John is a bachelor." If we are seriously interested in understanding the regularities in the use of real language in real situations we will be misled by persisting with idealizations of objective truth. As Lakoff (this volume) points out, ordinary language leans more to metaphor than to mathematics.

But there is a problem with unbridled relativism. If the "meaning" of an utterance can only be described in terms of its effects on a particular understander with a particular history, how do we talk about inter-subjective meaning at all? Since no two people have identical histories, and since any aspect of cognitive structure can potentially have an effect on the processing triggered by a particular utterance, there is a different meaning for every hearer. There is no objective "right meaning"—only a meaning for a particular person at a particular moment in a particular situation. Carried to an extreme, if you interpret my use of the word "dog" as referring to eggplant parmesan, what allows me to argue that you are wrong? We want to understand meaning in a way that makes sense of the fact that you and I may not have identical (or even mutually consistent) understandings of "democracy" or "system," but we cannot ignore the common sense intuition that there are broad areas of obvious agreement.

The most obvious fix is to take some kind of behaviorist or operationalist criterion. There is a long history in discussions of AI of using these criteria in arguing about whether a program can be said to "think" or "understand." The Turing test (Turing, 1950) is the most often cited form of an operational criterion, and many defenses of AI naively take for granted that it is the only sensible (scientific) way to deal with the philosophical issues. Even though no two individuals have the same internal structure, we can talk about equivalent classes of behavior and say that a person "understands" something if he behaves in the appropriate ways upon hearing it. But this is only sweeping the problem under a different rug. In trying to define what constitutes the class of "appropriate behavior on hearing the word 'dog' " we are stuck with a problem no easier than that of defining the meaning of "dog." If we posit some objective standards by which the behavior can be tested, we are making the same kinds of assumptions as in setting objective definitions for words. If we don't, then the question

"appropriate behavior according to whom in what situation" is just as troubling as "meaning according to whom in what situation." We cannot avoid relativism by converting it to objective propositions about behavior, and thus must deal in some other way with the problem.

The Problem of Representation

The other problem lies in the assumption that we can build formal structures that represent the knowledge and state of a language understander. In going from a standard deductive system to a structure-dependent formalism like KRL, we still maintain the basic idea of representing the relevant knowledge in formal structures that can be set down according to our understanding of linguistics, psychology, and computation. The question of whether human knowledge can be represented in formal structures has been a major concern for a number of philosophers. Dreyfus (1979) has been the most active critic of artificial intelligence from this perspective, drawing on the philosophy of Heidegger (1962).[18]

For a long time I found the arguments rather incomprehensible, seeing the position that knowledge was not representable as equivalent to the belief that the human nervous system could not operate according to deterministic principles. But Maturana, starting by viewing the nervous system as a mechanistic system, argued in ways that were disturbingly similar. One of the most challenging of Maturana's views is his dogmatic insistence that cognition is not based on the manipulation of mental models or representations of the world. For someone trained in AI (or in cognitive science generally, as illustrated by the other papers in this volume) it is hard to understand what other kind of explanation there could be.

Maturana sees much of the discussion of representation as exhibiting a serious error of confusing "phenomenic domains." Anything we choose to describe as a system can be described in different domains,[19] each with its relevant pheomena. For example, we can look at a TV screen and see an array of luminescent dots excited by a moving electron beam, or we can see a comedian telling jokes. We can talk coherently about what we see in either domain, but cannot combine them meaningfully. Maturana argues that in describing cognition we often fail to carefully distinguish the relevant domains. The error takes the form:

1. A scientist observes some recurrent pattern of interactions of an organism.
2. He or she devises some formal representation (for example a set of generative rules or a "schema") that characterizes the regularities.

[18]For a discussion of the issues from within AI, see Barr (1980).

[19]This use of the word "domain" follows Maturana. It is somewhat different from the more common use in AI, where the "domain" of a particular program is something like "blocks," "airline reservations," or "blood infections."

3. The organism is assumed to "have" the representation, in order to be able to exhibit the regularities.
4. (Depending on the particular sub-field) The scientist looks for experiments that will demonstrate the presence of the representation, or designs a computer program using it to see whether the behavior can be generated by the program.

The error is in the reification of the representation at step 3. Working from basic biological examples, Maturana points to many phenomena that *for an observer* can be described in terms of representation, but that can also be understood as the activity of a structure-determined system with no mechanism corresponding to a representation. As a simple example, we might watch a baby successfully getting milk from its mother's nipple and argue that it has a "representation" of the relevant anatomy, or of the activity of feeding. On the other hand, we might note that there is a reflex that causes it to react to a touch on the cheek by turning its head in that direction, and another that triggers sucking when something touches its mouth. From the viewpoint of effective behavior, it has a "correct representation," but it would be fruitless to look for neurophysiological mechanisms that correspond to reasoning that uses facts about breasts or milk.[20]

Maturana argues that there is a "domain of description" in which it is appropriate to talk about the correspondence between effective behavior and the structure of the environment or "medium" in which it takes place, but that we must not confuse this kind of description with the description of the causal mechanisms operating to produce it. In saying that a representation is "present in the nervous system," we are indulging in misplaced concreteness and can easily be led into fruitless quests for the corresponding mechanisms. Whereas the point is obvious for reflexive behavior (which can certainly be quite complex, as pointed out by the animal ethologists), he sees it as central to our understanding of all behavior, including complex cognitive and linguistic activities.

After an initial scepticism (of the "How could it be anything but . . ." variety), I thought about how this view might be directly applied to the problems of language understanding. There is a good deal of confusion of domains apparent in the work on "schemas," "scripts," and "frames." Some kind of regularity is observed in text patterns, or the ability to answer certain kinds of questions given text. The cognitive researcher builds a formal representation of this pattern, and often builds some kind of program that uses it to produce minor variants on the observed behavior. The resulting claim is that a person must "have" the script or schema and use it explicitly (perhaps not consciously) in carrying out the process.

[20]Geschwind (this volume) exhibits this kind of reification in most direct form: ". . . the cat who has never seen a mouse attacked will bite through the nape of the neck, thus adopting the "best strategy" for immobilizing the prey. These findings suggest the surprising conclusion that a "model" of the nape must be present in the nervous system . . ."

My own discussion of discourse (Winograd, 1977a) carried this representational view to its logical extreme. It gives examples of many different dimensions of discourse patterning, referring to them as kinds of "schemas." "Each speaker of a language possesses a large and rather diverse set of schemas dealing with the process of natural language communication. The understanding of these schemas will form the core of a science of discourse." I still feel that the kinds of phenomena that were pointed out and categorized were interesting and important, but dressing up the observations in the language of schemas did little or nothing to sharpen or develop them. A direct implementation of the purported schemas as representation structures in a computer program would have been uninteresting. It would have the flavor of much of the work we see, in which a program "plays back" the schemas or scripts put into it, but in doing so it does not provide any insight into what happens in those cases that don't closely match exactly one of the hypothesized schemas. *The schemas correspond to classes of external behavior, which may not correlate in any straightforward way to the components of the internal mechanism (either physical or functional).*

It was also interesting that the same questions could be raised with respect to computers. In other work (Winograd 1979b) I have been concerned with the problem of languages and programming environments for developing complex computer systems. The appropriate level of description for explaining a system to another person often includes terms that do not correspond to any mechanism in the program. If I say of a program "It has the goal of minimizing the number of jobs on the waiting queue," there is unlikely to be a "goal structure" somewhere in memory or a "problem solving" mechanism that uses strategies to achieve specified goals. There may be dozens or even hundreds of places throughout the code where specific actions are taken, the net effect of which is being described. In the case of computer systems (as opposed to naturally evolving biological systems) the goal can be more appropriately thought of as existing in the specification that was given to the people who programmed it. The situation would be more parallel to that of a living mind if the system had simply "survived" because a collection of changes led to a particular behavior pattern without conscious intention by programmers. But in any case, there is an important lesson in the fact that there are important regularities in the "descriptive domain" that *do not* correspond to mechanisms in the program.

Returning to the problems of language understanding, we can see what the appropriate domain is for talking about context. In traditional linguistics, the context of a sentence is made up of the sentences preceding and following it. In the AI model, as described in Figure 2 above, we assume that context can be understood as a set of cognitive structures within the speaker and hearer. Some of these structures are records (or analyses) of other sentences, but others are things like a "topic structure" representing what the discourse is about, a "script" that is being applied, and a "focus list" of things recently mentioned or thought about. Taking Maturana's examples seriously, we are led to question whether

these descriptive notions can appropriately be used as a guide for building or analyzing mechanisms. Perhaps some of the difficulties mentioned in section 3 result from an attempt to characterize regularities in the wrong domain. We need to ask, then, just what the relevant domain might be.

7. UNDERSTANDING IN A DOMAIN OF ACTION

Four Domains for Understanding Language

In getting to this point we have described language in three different phenomenic domains. The assumptions central to a given domain form a major element in the generation of a "paradigm" in the sense of Kuhn (1962). However, it is not a simple matter of choosing the "right" domain. For any field of inquiry, there can be several relevant domains of regularity. In focussing on one domain as central we are led to ask certain questions and pay attention to certain phenomena. For any choice, there are some phenomena that become more easily describable, and others that become more obscure. The three domains we have discussed so far are:

The domain of linguistic structure. This is the domain of traditional linguistics. One looks for regularities in the patterning of structural elements (phonemes, words, phrases, sentences, etc.) in utterances and text. As mentioned above, most of the work on the larger-scale structure of discourse is in this domain, even when it is reformulated in terms of "schemas."

The domain of correspondence between linguistic structures and the world. In this domain, one is concerned with regularities in the correspondence between the structures of linguistic objects and the states of affairs in the world that those objects describe. Much of the current work in the philosophy of language is an attempt to formalize this correspondence. Much of the AI work on natural language has had this orientation as well.

The domain of cognitive processes.[21] In this domain the relevant regularities are not in the linguistic structures themselves, or their correspondence to a world, but in the cognitive structures and processes of a person (or machine) that generates or interprets them. This is the domain explored in much of cognitive psychology and artificial intelligence.

My current work is moving in the direction of a fourth domain for understanding language:

The domain of human action and interaction. In this domain the relevant regularities are in the network of actions and interactions within a human society. An utterance is a linguistic *act* that has consequences for the participants, leading to other immediate actions and to commitments for future action.

[21]This view is used as a basis for a comprehensive description in a textbook entitled *Language as a Cognitive Process* (Winograd, forthcoming) and is contrasted with the first two views in Winograd (1977b).

This domain has been explored under the rubric of *speech acts*. The work of Austin (1962) and Searle (1970, 1975) has shown that utterances can be understood as acts rather than as representations. In giving a command or making a promise, a person is not uttering a sentence whose meaning lies in whether it does or does not correspond truthfully to the world. The speaker is entering into an interaction pattern, playing a certain role, committing both speaker and hearer to future actions, some linguistic and others not. In this domain the relevant question about an utterance is "What is the speaker doing?" Understanding is connected with the ability to recognize what the speaker is doing and to participate in the appropriate pattern of actions.

My first reaction on being urged to look at this domain was "Oh, just speech acts." I had been familiar with the basic ideas of speech acts for quite a while and viewed them as a kind of add-on to the central theory of language. In the paper on discourse discussed above they were described, true to form, as a collection of "speech act schemas" that were among the cognitive structures of a speaker or hearer. The challenge, though, was to see what it meant to look at language as a whole from the perspective of speech acts—as action rather than structure or the result of a cognitive process.

This was a difficult shift of view, but one with interesting consequences. It is perhaps easiest if we begin by looking at one of the most often described speech acts, that of *promising*. If someone asks me to come to a meeting tomorrow and I respond "I'll be there," I am performing a speech act[22] that is quite different from describing a state of affairs in the world or making a prediction about future states of affairs. By virtue of the utterance, I am entering into a commitment. It is not relevant to ask whether the promise is "true" or "false," but rather whether it is appropriate, or, to use Austin's term, "felicitous."

An essential thing about speech acts is that they always occur in a social context, with a background implied by that context. If I find out tomorrow that the meeting has been moved to Katmandu and don't show up, I can justifiably argue with you that I haven't broken my promise. What I really meant was "Assuming it is held as scheduled. . . ." On the other hand, if the meeting is moved to an adjacent room, and I know it but don't show up, you are justified in arguing that I have broken my promise, and that the "Katmandu excuse" doesn't apply. This kind of argumentation can be pursued back and forth indefinitely, and forms a part of the network of potential actions related to a promise. The legal system provides a regular mechanism for exactly this kind of interaction.

[22]As Searle points out, we can talk about a speech act as being a promise even though it does not use any explicit illocutionary verb, such as "promise" or "assure." Technically, it is said that the utterance has "promise" as its "illocutionary force."

Statements as Speech Acts Initiating Commitment

In the basic works on speech acts, there is a separation between the propositional content of an utterance and its illocutionary force. The fact that my utterance is a promise is its illocutionary force. The fact that it involves my attendance at a particular meeting at a particular time is its propositional content. In further pursuing the connection between meaning and speech acts, it is possible to view more of meaning (including what has been seen as propositional content) in the domain of action, rather than the domain of correspondence with the world.

Consider the following dialog:

A: I'm thirsty.

B: There's some water in the refrigerator.

A: Where? I don't see it.

B: In the cells of the eggplant.

A claims that B's first response was a lie (or "misleading"), whereas B contends that everything he said was literally true. Most work in semantics (including artificial intelligence) can be seen as providing formal grounds to support B. But there is an important sense in which a theory of "meaning" needs to deal with the grounds for A's complaint. In making the statement "There's some water in the refrigerator" B is doing something more than stating an abstract objective fact.

At first, it seems like it might be possible to expand on the definition of "water." Perhaps there is a "sense" of the word that means "water in its liquid phase in a sufficient quantity to act as a fluid," and the statement about water is ambiguous in whether it refers to this sense or to a sense dealing purely with chemical composition. But this doesn't help us in dealing with some other possible responses of B to the initial request:

B: There's no water in the refrigerator, but there's some lemonade.

B: There's a bottle of water in the refrigerator, with a little lemon in it to cover up the taste of the rust from the pipes.

In the first case, the presence of lemon in the water is taken as making it not "water." In the second, the lemon (perhaps the same quantity) is considered irrelevant. The difference lies in a background of assumptions the speaker has about the hearer's purposes and experience. After any amount of fiddling with the definition, one can always come up with a new context (e.g., what if the person were doing a science experiment or checking for sources of humidity), in which the acceptability of the statement "There is water" would not be accounted for by the definition. Every speech act occurs in a context, with a background understood by speaker and hearer. There are "felicity conditions"

that depend on mutual knowledge and intentions. The speaker is responsible for things he can anticipate that the hearer will infer from what he says, not just its abstract correspondence with the state of affairs.

What happens, then, if we try to understand the problem of "truth" in the terms of social action and commitment? In making a statement I am doing something like making a promise—committing myself to acting in appropriate ways in the future. In this case, there is a different kind of satisfaction condition. There is no specific action that I am bound to, but there is a structure of potential dialog that we could enter into in the face of a "breakdown." If I say "There is water in the refrigerator" and you can't find any, I am committed to give you an account. Either we agree that I was wrong, or we discuss the assumed background ("I assumed we were talking about something to drink." "I assumed we were talking about chemical composition").

There are several reasons why this shift of viewpoint is potentially advantageous:

It lets us deal with what happens when we actually make statements. Formal approaches to meaning often take as their model the language of mathematics, in which it is generally assumed that the truth of a statement can be determined without reference to outside context or situation.[23] But in real language, we rarely if ever make a statement that could not be construed as having a literal meaning we don't intend. If I say "snow is white" you can point to the murky grey polluted stuff at our feet. I reply "I meant pure snow," and you respond "You didn't say so, and anyway no snow is absolutely pure." It is an edifying exercise to look at the statements we make both in our writing and our everyday conversation and see how few of them can even apparently be judged true or false without an appeal to an unstated background.

It shifts us out of the objective/subjective dichotomy. In the previous section we saw a dilemma arising from trying to identify the "meaning" of a word or utterance. By assuming it had an objective meaning independent of a particular speaker/hearer situation, we banished many aspects of meaning that play a central role in language. But in assuming that meaning is defined in terms of effect on a particular individual in a situation, we lose the sense that meaning can be the same across individuals and situations. In moving to the domain of interactions (rather than that of objective truth correspondence or cognitive process), we are directing attention to the interactional situation in which something is uttered. We draw generalizations across these situations (and their potential for continued conversation) rather than across objective correspondence or mental states.

It places central emphasis on the potential for further articulation of unstated background. By looking at statements as language acts analogous to promises, we bring into prominence the fact that human action always occurs in an unarticulated background. When I promise to do something, it goes without saying that the commitment is relative to a large number of assumptions about the rest of our world continuing as expected. The same properties carry over to language acts. Sociologists working in "ethnomethodology"[24] have explored the problems in recognizing our background

[23]In fact, as pointed out by Lakatos (1976), this is not really the case even in mathematics.
[24]For example, Garfinkel (1967).

assumptions. We can never make the background fully explicit, but we can study the nature of the dialog by which people come to a consensus about things that were previously in the background.

8. WHITHER FROM HERE?

Having turned to a different "phenomenic domain" we must raise the question of methodology. How does one proceed in day to day work? My own answer to this question includes both constructive and interpretive activities:

Developing Theories and Formalisms

Just as the structural formalisms of linguistics and the deductive formalisms of predicate calculus were developed to provide a language for description in their respective domains, we need an appropriate "calculus of language acts" if we want to develop detailed theories of language interaction. There will be several parts of such a theory:

1. *Illocutionary logic.* The basic groundwork of speech act theory includes an analysis of the different kinds of illocutionary points and the structure of the felicity conditions associated with them. Searle (1975) proposes a universal taxonomy of five basic illocutionary points, further developed by Searle and vanDerVeken (in press). This analysis can serve as a starting point for understanding the structure of larger composite patterns made up of the basic acts. For example an "offer-negotiation-acceptance" sequence is a standardly observed pattern made up of individual "commissives" and "requests." The formal tools for describing the "syntax" of such patterns may be quite different from those used in traditional linguistics, since they must take into account the passage of time (e.g., not overtly responding to an invitation constitutes a kind of response).

2. *Taxonomy of linguistic grounding.* In order to carry out the suggested program of viewing the truthfulness of statements in the domain of interaction, we need a "logic of argument," where "argument" stands for the kind of elucidation of background assumptions discussed above. When I make a statement, I am making a commitment to provide some kind of "grounding" in case of a "breakdown." This grounding is in the form of another speech act (also in a situational context) that will satisfy the hearer that the objection is met. There appear to be three basic kinds of grounding: experiential, formal, and social.

Experiential: If asked to justify the statement "Snow is white" I may give a set of instructions ("Go outside and look!") such that a person who follows them will be led to concur on the basis of experience. The methodology of science is designed to provide this kind of grounding for all empirical statements. Maturana (in press) points out that the "objectivity" of science derives from the assumption that for any observation, one can provide instructions that if followed by a "standard observer" will lead him or her to the same conclusion. This does not necessarily mean that the result is observer free, simply that it is anticipated to be uniform for all potential observers.

Formal: Deductive logic is based on the playing of a kind of "language game" in which a set of basic rules are taken for granted, and argument proceeds as a series of moves constrained by those rules. For example, if I expect you to believe that all Swedes are blonde and that Sven is a redhead, then I can use a particular series of moves to provide grounding for the statement that Sven is not Swedish. Of course, this depends on the grounding of the statements used in the process—one can recursively demand grounding for each of them. Under this category fall most of the issues that have been discussed in formal semantics,[25] but with a different emphasis. The focus is not on their coherence as a mathematical abstraction, but on the way they play a role in the logic of conversation.

Social: Much of what we say in conversation is based neither on experience nor logic, but on other conversations. We believe that water is H_2O and that Napoleon was the Emperor of France not because we have relevant experience but because someone told us. One possible form of grounding is to "pass the buck"—to argue that whoever made the statement could have provided grounding. This is also recursive, but we assume that the buck stops somewhere. Of course this need not be so (as illustrated by the development of ideologies in societies), but this is not relevant to its role in the dynamics of conversation.

Just as one can develop taxonomies and structural analyses of illocutionary points, it is important to develop a precise analysis of these structures of argumentation. There are many ways in which such a logic will parallel standard formal logic, and others in which it will not. In particular, it seems that the role of analogy and metaphor will be much more central when the focus is on patterns of argumentation between individuals with a shared background rather than on deductive inference from axioms.

In thinking about these problems it is interesting to look at the structure of reasoning done by lawyers rather than mathematicians. The lawyer is not guided by a formal theory of proof that can be used mechanically to establish the validity of a conclusion. Instead, he or she attempts to anticipate the objections that could be raised (by another person in a particular context) and to prepare justifications that can overcome the objections. Those justifications are in turn based on statements that may themselves need other justifications, *ad infinitum*. Superficially, there is an analogy to the hierarchical structure of theorems based on simpler theorems in mathematics, but there is a deep fundamental difference. The legal brief does not "ground out" on unquestionable axions. Instead it rests on suppositions that the lawyer assumes are part of the basic background (experiential, social, etc.) and hence will not be called into question. Of course this is only an anticipation, and there are mechanisms in court for going deeper into the logic of any statement in the form of an adversary dialog.

3. *Theory of reasoning as triggered by breaking-down.* Going one step further, we need to inquire into the relationship between argumentation and reasoning. The classical understanding is that reasoning proceeds by logical

[25]Hintikka (1976) uses the notion of games in his development of deductive logic, including modal logic.

deductive steps, and argumentation is a laying-out of the steps for someone else to follow. The orientation provided here is almost the reverse—reasoning is a form of arguing with yourself, and the justification for a step is that it (temporarily, at least) quiets objections to some statement. The motivation for going farther is that something that previously seemed obvious (either on examination or because it had never been examined) has broken down. Starting with the analyses of philosophers like Heidegger (1962) we can look for a "logic of breaking down" that applies not only to linguistic conversations, but to our non-verbal interactions and understanding.

There are interesting ways in which its characteristics (of breakdowns leading to questions to a certain depth, still within an implicit background) are parallel to those of systems using resource dependent reasoning. Perhaps the mechanisms of systems like KRL can be usefully reinterpreted as representing a logic of argument, rather than a logic of cognitive processing. The issues raised in these systems (such as default knowledge, meta-description, and partial matching) have intriguing connections to problems of background and representation. It would be presumptuous to see the existing formalism as an "answer," but it provides a starting point for exploring the interaction between implicit background, formal structures, and patterns of conversation and justification.

Understanding Existing Systems and Designing New Ones

Computer Systems. Much of cognitive science has been both stimulated by and directed towards the construction of computer programs that behave "intelligently." Hundreds of books and articles have been written on how computer systems will soon become prevalent in every arena of life. The question asked of the cognitive scientist is: "What kind of theories do we need in order to build intelligent systems we can use?"

The prevalent view is that in AI we design "expert systems" that can stand as surrogates for a person doing some job. From a viewpoint of human interaction we see the computer's role differently It is not a surrogate expert, but an intermediary—a sophisticated medium of communication. A group of people (typically including both computer specialists and experts in the subject domain) build a program incorporating a formal representation of their beliefs. The computer communicates their statements to users of the system, typically doing some combinations and rearrangements along the way. The fact that these combinations may involve complex deductive logic, heuristic rule application or statistical analysis does not alter the basic structure of communicative acts.

A person writing a program (or contributing to its "knowledge base") does so within a background of assumptions about how the program will be used and who will be interpreting its responses. Part of this can be made explicit in

documentation, but part is an implicit background of what can be "normally understood." Except for systems operating within strongly constrained domains, there inevitably comes a time when the system "breaks down" because it is being used in a way that does not fit the assumptions. This is true not only of "expert systems," but of computer programs in all areas. Many of the biggest failures of mundane computer systems (management systems, inventory systems, etc.) have come not because the system failed to do what the designers specified, but because the assumptions underlying that specification were not appropriate for the situation in which the program was used. This will become even more the case as we build systems that are more flexible—that allow the user to develop new modes of interaction in the course of using the program, rather than staying within a fixed set of alternatives.

Of course, to the degree that system builders can anticipate areas of potential breakdown, they can make explicit statements in advance, which the computer can convey (again perhaps with complex recombination) through an "explanation system."[26] Some existing programs incorporate explanation facilities that move in this direction, but they are able to deal with only a limited range of the potential dialogs of explanation. There is always a limit set by what has been made explicit, and always the potential of breakdowns that call for moving beyond this limit.

If we see the machine as an intermediary, it is clear that the commitment (in the sense discussed above in viewing truth in the context of speech acts) is made by those who produce the system. A dialog must be carried on with the people who performed the original speech acts, or those to whom they have delegated the responsibility. In the absence of this perspective it becomes easy to make the dangerous mistake of interpreting the machine as making commitments, and losing the sense that some person or group of people has responsibility for what it does. We need theories that can help in understanding the properties of dialogs in which a person tries to elucidate the background assumptions that may have led to a breakdown. In designing the computer system the focus is shifted from the problem of "How do we make sure it always gives the right answer?" to "What is the best kind of organization for figuring out when and where things are going wrong?"

Medical diagnosis programs provide a good example. Imagine a program written by a team of computer specialists, working with a group of medical experts. It is installed by the administration of a hospital and used by a member of the medical house staff in choosing a treatment. If the diagnosis was wrong and the patient is harmed, who is responsible? The problem may not be one of wrong medical knowledge, but rather one of background assumptions. An answer that is correct in one context may be inappropriate in another. For example, if the program was written with ambulatory patients in mind, it might not be

[26]Winograd (1979a) discusses the importance of developing systems of this type.

appropriate for a chronic bedridden invalid who shows very different symptom patterns. How can the user of the system systematically find out what the relevant assumptions are? If we do not develop theories that help us to cope with these issues, the systems will be useless or worse.

I also believe that there are potential new directions in the design of computer languages in which these issues will play a prominent role. The writing of a program is a communication between one programmer and another. The understanding of existing programs is one of the most important and difficult things we do, and future systems will need to be oriented primarily towards aiding in this understanding. The issues of language and understanding that have been at the heart of this paper will be at the heart of those systems.[27]

Management. Another practical area that looks to cognitive science for a theoretical base is management. There is a significant overlap of concepts, as illustrated by terms like "problem solving," "decision making," and "heuristic strategy." Of course this is no accident. Herbert Simon, one of the founders of cognitive science, is best known for his work on organizational decision making. The management and decision metaphor has played a large role in shaping artificial intelligence from the earliest days.

My own experience is not in this area, but my work with Fernando Flores (who has been a cabinet minister in the Chilean government) has pointed out concerns that are parallel to those arising in the study of language. The view of management as problem solving and decision making suffers from the same problems of background and context. In looking at the role of a manager as optimizing some value by choosing among alternatives for action, we are in the same position as in looking at the understanding of a word as choosing among a set of formal definitions. The hard part is understanding how the alternatives relevant to a given context come into being. The critical act of problem solving in the recognition of the problem.

Again, this is not a new insight. Management specialists (including Simon) have at times pointed out the essential role of "problem acquisition" as a prerequisite to "problem solving." The practical wisdom of management recognizes the importance of the role a manager plays in continually creating the "problem space" in which he or she operates. But the theories provided by cognitive science and artificial intelligence have little to say about this process of creation. To a large extent, a problem is created by the linguistic acts in which it is identified and categorized. Of course, some situation exists previous to the formulation, but its structure as a problem (which constrains the space of possible solutions) is generated by the commitment of those who talk about it.

Once again we cannot look to simple notions of truth and deduction. The "energy crisis" was not created by specific acts of the oil companies, the Arabs,

[27]Winograd (1979b) discusses these issues in a more technical vein.

or the American consumer, but by those with the power to create consensus who looked at a long-term situation and determined it to be a crisis. The relevant question is not whether it is "true" or "false" that there is a crisis, but what commitments are entailed (for speaker and hearer) by the speech acts that created it.

Education. A third area where cognitive theories play a central role is in education. To a large extent our current educational practice has grown from experience with no theory at all, or with a hodgepodge of theories borrowed from behaviorism, structuralism, and various kinds of interpersonal psychology. There have been attempts to attract educators to a theory based on work in artificial intelligence, emphasizing how human thought can be described in terms of concepts such as "algorithm," "heuristic," and "bug." Researchers like Papert (1972) point out that in teaching children subject material we are also "teaching children thinking." In providing instructions for how to go about particular tasks we are also providing models for how to go about tasks in general. By being conscious of this "meta-teaching" and by applying the best theories we can of thought and language, we can provide a more consistent and effective educational environment.

Again, the questions we choose to look at depend on the domain we see as central. If we concentrate on cognitive processing, students will learn to think in that domain. Faced, for example, with a breakdown in some kind of interpersonal situation, they will ask "What is the bug in my model of the other person?" Starting from an orientation toward the structure of interaction, the question becomes "What do we do to make explicit the assumptions that led to the breakdown?" The emphasis is on the nature of the continuing interaction, rather than on the cognitive structure of the participants. No one orientation is good for everything, and the purpose of an educational system should be to give students experience with the broadest possible range. There is a significant place for understanding language and thought in terms of social interaction.

9. CONCLUSION

Those who are familiar with the philosophical discussion about artificial intelligence will have noticed that many of the ideas and sources discussed here (such as the incompleteness of formal representation, Heidegger's notion of background, etc.) are among those cited by critics like Dreyfus (1979), who deny the possibility of developing any formalization of human thought and knowledge. A conclusion one might draw is that having brought these questions to light, we can only proceed by abandoning formal cognitive science and our attempts to program computers to do things we consider "intelligent."

It should be clear from the previous section that this is not my conclusion. I am not advocating (or planning) the abandonment of the scientific study of cognition, but trying to better understand what we are doing, and refocusing my efforts in the light of that understanding. However, it could be argued that the path described in this paper is one leading away from the paradigm of cognitive science. Even granting that an orientation towards language as social action is interesting and useful, it is arguable that it is not "cognitive science"—that it represents a turning away from the domain of "cognitive processing" (or, as it is often called, "information processing"). In some ways this observation is valid, but in others it is misleading.

It is important to recognize what we are doing we apply words like "cognitive" or "science" to a particular enterprise or approach. In our writing, teaching, and interactions with people (both in the field and outside), we are performing speech acts that give those words meaning. Different orientations lead to different practical suggestions, to different ways of interpreting and acting. As has been pointed out in the philosophy of science, the importance of a paradigm may not lie so much in the answers it provides as in the questions it leads one to consider, and a paradigm (like a crisis) is created by a web of interlinked speech acts.

Some day in the future there may be a field called "cognitive science" whose boundaries are defined by a narrow common approach and domain, just as there are fields of "anatomy," "physiology," and "pathology" dealing with the physical structures and functioning of the human body in their own domains. This narrowly defined cognitive science would deal with those aspects of language, thought, and action that are best understood in the domain of information processing. At the moment though, this is not the case. As indicated by this volume and the nascent professional society it represents, "cognitive science" is a broad rubric, intended to include anyone who is concerned with phenomena related to mind. I believe that the kinds of issues I have been led to in looking at language are relevant to a broad segment of cognitive science as generally interpreted, and a serious consideration of them may call into question many of the assumptions at the heart of our current understanding.

ACKNOWLEDGMENTS

My thinking about language has been developed through interactions with a number of people, as chronicled in the paper. I am particularly indebted to Fernando Flores for continually introducing new vistas. It was in conversations with Danny Bobrow and Brian Smith that the ideas of KRL emerged, and work on KRL has involved many other people. Coversations with Don Norman led to some of the original thinking about memory structure. Henry Thompson, Rich Fikes, David Levy, Mitch Model, Paul Martin, Jonathan King, and Wendy Lehnert have all contributed to the ideas and implementations. In addition to Bobrow, Flores, Levy, and Thompson, the Berkeley discussions mentioned at the beginning of section 5 included Hubert Dreyfus, John Searle, and Stephen White. Danny Bobrow, Fernando Flores, David Levy, and David Lowe provided insightful comments on an earlier draft of this paper.

REFERENCES

Austin, J. L. *How to do things with words*. Cambridge, Mass.: Harvard University Press, 1962.

Barr, A. The representation hypothesis. (Stanford HPP-Memo, HPP-80-1, Dept. of Computer Science 1980).

Bobrow, D. G., & Winograd, T. An overview of KRL, a knowledge representation language. *Cognitive Science*, 1977, *1*, 3–46.

Bobrow, D. G., & Winograd, T. KRL: Another perspective. *Cognitive Science*, 1979, *3*, pp. 29–42.

Dreyfus, H. *What computers can't do: A critique of artificial reason*. (2nd Ed.) San Francisco: Freeman, 1979.

Feigenbaum, E., & Feldman, J. *Computers and thought*. New York: McGraw Hill, 1963.

Fillmore, C. An alternative to checklist theories of meaning. In Cogen et al. (Eds.), *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society*. University of California, Berkeley, 1975.

Flores, C. F., & Winograd, T. *Understanding cognition as understanding*. Norwood, N.J.: Ablex Publishing Corporation, in press.

Fodor, J. A. Tom Swift and his procedural grandmother. *Cognition*, 1978, *6*, pp. 229–247.

Fodor, J. A. Methodological solipsism as a research strategy in psychology. *Brain and Behavioral Sciences*, forthcoming.

Gadamer, H-G. *Philosophical hermeneutics*. Translated by David E. Linge. Berkeley: University of California Press, 1976.

Garfinkel, H. What is ethnomethodology? In H. Garfinkel (Ed.), *Studies in Ethnomethodology*. Englewood Cliffs, N.J.: Prentice-Hall, 1967.

Geschwind, N. Neurological knowledge and complex behaviors. *Cognitive Science*, 1980, *4*, pp. 185–193.

Grosz, B. J. Focusing and description in natural language dialogues. In A. K. Joshi, I. A. Sag, & B. L. Webber (Eds.), *Elements of discourse understanding*. Cambridge: Cambridge University Press, 1980.

Heidegger, M. *Being and time*. New York: Harper & Row, 1962.

Hintikka, J. Quantifiers in logic and quantifiers in natural languages. In S. Korner (Ed.), *Philosophy of Logic*. Berkeley, Calif: University of California Press, 1976.

Hobbs, J. R. Resolving pronoun references. *Lingua*, 1978, *44*, 311–338.

Katz, J. J., & Fodor, J. A. The structure of a semantic theory. In J. Fodor & J. Katz, (Eds.), *The Structure of Language*. Englewood Cliffs, N.J.: Prentice-Hall, 1964.

Kuhn, T. *The structure of scientific revolutions*. Chicago: Ill.: University of Chicago Press, 1962.

Labov, W. The boundaries of words and their meanings. In C-J. N. Bailey & R. Shuy (Eds.), *New ways of analyzing variation in English*. Washington, D.C.: Georgetown University, 1973.

Lakatos, I. *Proofs and refutations*. Cambridge, Mass.: Cambridge University Press, 1976.

Lakoff, G. & Johnson, M. The metaphorical structure of the human conceptual system. *Cognitive Science*, 1980, *4*, pp. 195–208.

Lighthill, Sir. J. *Artificial intelligence: A general survey*. London: Science Research Council, 1973.

Maturana, H. R. Biology of language. In R. W. Rieber (Ed.), *The neuropsychology of language*. New York: Plenum Press, 1977.

Minsky, M. *Semantic information processing*. Cambridge, Mass.: MIT Press, 1967.

Minsky, M. A framework for representing knowledge. In P. Winston, (Ed.), *The psychology of computer vision*. New York: McGraw-Hill, 1975.

Minsky, M. K-Lines: A theory of memory. *Cognitive Science*, 1980, *4*, pp. 117–133.

Norman, D. A. Twelve issues for Cognitive Science. *Cognitive Science*, 1980, *4*, pp. 1–32.

Palmer, R. E. *Hermeneutics: Interpretation theory in Schleiermacher, Dilthey, Heidegger and Gadamer*. Evanston, Ill.: Northwestern University Press, 1969.

Papert, S. Teaching children thinking. *Mathematics Teaching: The Bulletin of the Association of Mathematics* No. 58, Spring, 1972.

Rosch, E. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 1975, *104*, pp. 192–233.

Schank, R. & Abelson, R. *Scripts, plans, goals, and understanding*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1975.

Schank, R. C. Language and memory. *Cognitive Science*, 1980, *4*, pp. 243–284.

Searle, J. *Speech Arts*. Cambridge, Mass.: Cambridge University Press, 1970.

Searle, J. A taxonomy of illocutionary acts. In K. Gunderson (Ed.), *Language, mind, and knowledge: Minnesota studies in the philosophy of science*, Vol. XI. Minneapolis, Minn.: University of Minnesota Press, 1975.

Searle, J. R. The intentionality of intention and action. *Cognitive Science*, 1980, *4*, 47–70.

Turing, A. M. Computing machinery and the mind. *Mind: A Quarterly Review of Psychology and Philosophy*, 1950.

Winograd, T. *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language*. MAC-TR-84, MIT Project MAC, 1971.

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

Winograd, T. A process model of language understanding. In R. C. Schank & K. M. Colby (Eds.), *Computer models of thought and language*. San Francisco, Calif.: W. H. Freeman, 1973.

Winograd, T. When will computers understand people? *Psychology Today*, 1974, *7*, pp. 73–79. (a)

Winograd, T. Five lectures on artificial intelligence, PIPS-R. No. 5, Electrotechnical Laboratory, Tokyo, Japan, 1974. Reprinted in A. Zampolli (Ed.), *Linguistic structures processing*. North Holland, 1977. (b)

Winograd, T. Frame representations and the procedural–declarative controversy. In D. Bobrow & A. Collins (Eds.), *Representation and understanding: Studies in Cognitive Science*. New York: Academic Press, 1975.

Winograd, T. Towards a procedural understanding of semantics. *Revue Internationale de Philosophie*, 1976, *3*, pp. 117–118.

Winograd, T. A framework for understanding discourse. In M. Just & P. Carpenter (Eds.), *Cognitive processes in comprehension*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977. (a)

Winograd, T. On some contested suppositions of generative linguistics about the scientific study of language. *Cognition*, 1977, *5*, pp. 151–179. (b)

Winograd, T. On primitives, prototypes, and other semantic anomalies. *Proceedings from Theoretical Issues in Natural Language Processing II*. University of Illinois at Champaign–Urbana, 25–32, 1978.

Winograd, T. Towards convivial computing. In M. L. Dertouzos & J. Moses (Eds.), *The computer age: A twenty-year view*. Cambridge, Mass.: MIT Press, 1979. (a)

Winograd, T. Beyond Programming languages. *Communications of the ACM*, 1979, *22*, 391–401. (b)

Winograd, T. Extended inference modes in computer reasoning systems. *Artificial Intelligence*.

Winograd, T. *Language as a cognitive process*. Reading, Mass.: Addison–Wesley, (in preparation).