Some Methodological Issues in
Natural Language Understanding Research

W. A. Woods
Bolt Beranek and Newman, Inc.
Cambridge MA 02138

I. INTRODUCTION

Natural language understanding has suddenly emerged as a central focus for many different disciplines. Applications are emerging all over the field of computer science in which language understanding and the communication of complex intentions to machines is a crucial part. Moreover, psychologists, linguists, and philosophers have found the models emerging from computational linguistics research to provide new stimulus and new methods for increasing our understanding of the process of human language use and the nature of communication. In this paper I want to discuss some of the methodological problems I see in the development of this area of research and some of the things which I think are needed in order for the field to be productive of real scientific insight and useful results.

In order to discuss methodologies, we had best first understand the different tasks for which the methodologies are to be used. There are at least two primary interests which one can have in studying natural language understanding -- constructing intelligent machines and understanding human language performance. These two different objectives are not mutually exclusive, and I will attempt to argue that a large portion of the research necessary to either of them is shared by the other. This common portion consists of a pure attempt to understand the process of language understanding, independent of what device (human or machine) does the understanding. However, there are elements of the different points of view which are not shared, and drawing the distinction between objectives at the outset is, I think, useful.

I would claim that both the development of useful mechanical devices for understanding language and the understanding of human language performance depend heavily on what we might call "device independent" language understanding theory. That is, a joint study of human and machine language understanding, attempting to devise algorithms and system organizations which will have the functional performance of language understanding without specifically trying to model the performance aspects of human beings. Theoretical and empirical studies of this sort provide the foundations on which models of human language processing are built which are then subject to empirical verification. They also provide the "bag of tricks" out of which useful mechanical language understanding systems can be constructed. Outside the common area of endeavor, these two different objectives have different goals. For both objectives, however, a major component of the research should be to study the device independent language understanding problem.

This paper is an attempt to set down my biases on some issues of methodology for constructing natural language understanding systems and for performing research in computational linguistics and language understanding. In it I will discuss some of the methods that I have found either effective and/or needed for performing useful work in the area of human and mechanical language understanding.

For theoretical studies, I will argue strongly for a methodology which stresses communicable and comprehensible theories, with precise uses of terms and an evaluation of formalisms which stresses the cognitive efficiency of the representations of the formalism itself. I will attempt to cite several examples of the differences in cognitive efficiency between formalisms.

The thrust of many of my comments will deal with the problems of complexity. My thesis is that natural language, unlike many physical systems is complex in that it takes a large number of facts, rules, or what have you to characterize its behavior rather than a small number of equations (of whatever theoretical sophistication or depth). It is relatively easy to construct a grammar or other characterization for a fairly small subset of the language (at least it is becoming more and more so today), but it is not so easy to cope with the complexity of the specification when one begins to put in the magnitude of facts of language which are necessary to deal with a significant fraction of human language performance. Theories for natural language understanding will have to deal effectively with problems of scale the number of facts embodied in the theory.

Since this paper is largely designed to promote discussion, the set of issues covered herein makes no effort to be complete. My goal is to raise some issues for consideration and debate.

II. A PROGRAM FOR THEORETICAL LINGUISTICS AND PSYCHOLOGY

The first point that I would like to make is that in the pursuit of theoretical understanding in linguistics or psycholinguistics, studies will be much more productive if pursued in the context of total language understanding systems and not in isolation. The subdivision of the total process into components such as syntax and semantics and concentrating on one such component is an effective way of limiting scope. However, it is only justifiable if one has at least some reason to believe that his hypothesized interfaces to those other components are realistic (and certainly only if he has precisely specified those interfaces). One cannot expect to pursue some small niche of the language understanding process without an active interest in the entire process and an understanding of the role of his specialty area in that overall process. Otherwise it is too easy to push problems off on someone

else, who may not be there to catch them. (In particular there is considerable risk that the problems left for someone else may be insoluble due to a false assumption about the overall organization. Studies pursued under such false assumptions are likely to turn out worthless.)

III. THEORETICAL AND EMPIRICAL METHODOLOGIES

There is need in the field of natural language understanding for both theoreticians and builders of systems. However, neither can pursue their ends in isolation. As in many other fields, the theoretical and experimental components go hand in hand in advancing the understanding of the problem. In the case of language understanding, the theoretical investigations consist largely of formulation of frameworks and systems for expressing language understanding rules or facts of language and for expressing other types of knowledge which impact the understanding process. On the experimental side, it is necessary to take a theory which may appear beautifully consistent and logically adequate in its abstract consideration, and verify that when faced with the practical reality of implementing a significant portion of the facts of language, the formalism is capable of expressing all the facts and is not too cumbersome or inefficient for practicality. The day is past when one could devise a new grammar formalism, write a few examples in it, and tout its advantages without putting it to the test of real use.

Today's language theoreticians must have a concrete appreciation of the mechanisms used by computerized language understanding systems and not merely training in a classical school of linguistics or philosophy. (On the other hand, they should not be ignorant of linguistics and philosophy either.) Some mechanism must be found for increasing the "bag of tricks" of the people who formulate such theories -- including people outside the current computational linguistics and artificial intelligence camps. Hopefully, this conference will make a beginning in this direction.

IV. MODELS AND FORMALISMS

One of the depressing methodological problems that currently faces the field of artificial intelligence and computational linguistics is a general tendency to use terms imprecisely and for many people to use the same term for different things and different terms for the same thing. This tendency greatly hampers communication of theories and results among researchers.

One particular imprecision of terms that I would like to mention here is a confusion that frequently arises about models.

One frequently hears people refer to the transformational grammar model, or the augmented transition network grammar model, and asks what predictions these models make that can be empirically verified. However, when one looks carefully at what is being referred to as a model in these cases, we find not a model, but rather a formalism in which any of a number of models (or theories) can be expressed. The transformational grammar formalism and the ATN formalism may suggest hypotheses which can be tested, but it is only the attachment of some behavioral significance to some aspect of the formalism which gives rise to a testable model.

Arguments for or against a model are whether it is true -- i.e. whether the predictions of the model are borne out by experiments. Arguments for or against a formalism or a methodology are its productiveness, economy of expression, suggestiveness of good models, ease of incorporating new features necessary to new hypothesized models (i. e. range of possible models expressible), etc. One needs at the very least that the formalism used must be capable of representing the correct model. But one doesn't know ahead of time and may never know what the correct model is. Hence it is desirable to have a formalism that can represent all conceivable models that could be correct. If there is a class of models which the formalism cannot account for then there should be an argument that no members of that class could possibly be correct, otherwise a formalism which included that class would be better (in one dimension). Dimensions of goodness of formalisms include range of possible models, efficiency of expression (perspicuity or cognitive efficiency of the formalism), existence of efficient simulators for the formalism for use in verifying the correctness of a model, or for finding inadequacies of a model, or for determining predictions of the model, etc.

V. HUMAN LANGUAGE PERFORMANCE

In order to perform good work in computational linguistics and in understanding human language performance, one needs to keep always in mind a good overview of how people use language and for what. Indeed, a prime focus of this conference is the development of such a overview. My own picture of the role of language in human behavior goes roughly like this:

There is some internal representation of knowledge of the world which is prelinguistic, and we probably share most of it with the other higher animals -- I would guess we share a lot of it with cats and dogs, and certainly with apes and chimpanzees. (What differences of quality or quantity set us apart from these animals or set the chimps apart from the dogs I would not care to speculate.) Nevertheless, it is clear that cats and dogs without our linguistic machinery and without spoken languages do manage to store and remember and use fairly complex pieces of knowledge of the world, such as how to open

ors, how to find their way around, where
ey left things, which dish is theirs, what
nny sequence of sounds their owners use to
ll them (i. e. their names), and the
gnificance (to them) of all sorts of
ings that go on in the world around them.
mans probably remember large numbers of
ch things also without specific need for
nguage. We presumably have in our head
mething which is like a language in many
spects, but which probably does not share
e peculiar linear characteristics of
oken and written language (which derives
om the temporal order imposed on speech
d reading).

It no doubt helps us to remember a
rger number of things to correlate them
th linguistic labels or a pronounceable
quence of sounds, and this no doubt gives
 greater ability for abstract thought.
wever, it is doubtful that a language as
 speak it or write it is a prerequisite
r an organism to have what we might call
ought. Many of the things which we "know"
e not expressed in language, and the fact
at finding the appropriate words to
scribe things that we understand is
metimes very difficult should give us a
ue that the representation which we use in
r heads is not a simple transcription of
e language that we use to communicate with
hers. Rather, there are a variety of
position problems which need to be solved
 order to translate even ideas which are
early understood into a linear sequence of
nguistic symbols which will be likely to
ouse or create the same idea in the head
 our listener. It seems likely then that
e notation or whatever conventions that we
e to store ideas and information in our
ads is not the same as the language that
 speak to communicate with others.

The language that we speak and write,
en, appears to be a device or a discipline
olved for the purpose of attempting to
ouse in the head of the listener something
milar to that which is encoded in the head
 the speaker.

The process of communication
cessarily involves elements of problem
lving. What terms does my listener know?
at concepts can I rely on his
derstanding so that I can express what I
nt to say in terms of them? How can I
ild a specification out of these pieces
ich will cause him to construct in his
mory the thing which I am trying to
mmunicate? An account of human language
e must deal with all of these questions.

The above picture of the overall role
' language in human communication may not
 correct in many respects. Hopefully a
nsensus of this workshop will produce a
tter one. However, I am afraid that a
mplete understanding of human language use
ll have to go hand in hand with an
derstanding of the prelinguistic
pabilities for knowledge representation
d use which the human has. This level of
man ability is unfortunately very
fficult to get one's hands on since it,
ke Joe Becker's problems of intermediate

cognition, is a process which we are not
generally aware of (since it takes place
below the level of our conscious awareness)
and consequently we have no direct abilities
to see it. Rather we have to be able to
infer its presence and its nature from
theoretical considerations and the effects
that it has on the overt behavior we can
see. A methodology for working in this area
is extremely difficult to work out.

A principal component of such a
methodology, I feel, should be a theoretical
attempt to construct models which do things
humans do and which do them well. That is,
one should try to design intelligent
machines which can do what humans do, and
let the concepts that emerge from such
designs make predictions about what
performance one should see at the overt
behavior interface. It is important
however, that such studies go as far as to
produce overt behavior which can be
evaluated. A so called "theoretical" study
which has no measurable performance is
foundationless. There is no way to evaluate
whether it is doing anything or not. In
particular, many studies of so called
"semantic representations" need clear
statements of what will be done with their
representations and how one can decide
whether a representation is correct or
incorrect. Without such an understanding,
the entire exercise is one of aesthetics and
without scientific contribution. In talking
about semantic representations, one must be
willing to face the questions of how the
device knows what those representations
mean. What events in the world will be in
contradiction to the knowledge encoded in
the representation and what ones will be
consistent with it? How will a person (or a
machine) know whether an event perceived is
consistent with his semantic representations
or not? How does he decide what to record
when he perceives an event -- i. e. what
process transforms ("transforms" is not
really the right word for this) an observed
event into a linguistic description of it?
What intervening processes take place? These
and similar questions must be specifically
faced.

VI. EXPLANATORY MODELS

The goal in trying to model human
behavior should be to find explanatory
models, not just descriptive models. If,
for example, one discovers that there is a
reaction time lag in processing certain
types of sentences, a model which simulated
this behavior by inserting a delay into a
certain stage of the processing would be a
descriptive model, whereas another model
which took longer for processing these types
of sentences because of some extra
processing which they required due to the
organization of the program would be an
explanatory model. In my own work, the
things which have excited me and made me
feel that I was discovering something about
the way the people understand language, have
been algorithms that are motivated by
considerations of efficiency and "good
engineering design" for a specific task

which then turn out to have predictions which are borne out in human performance. An example of this is some of the work of Ron Kaplan and Eric Wanner using ATN grammars to model aspects of human linguistic processing. (The basic ATN grammar formalism was designed for efficiency of operation, and not specifically for human performance modeling.) When such an experiment has positive results, one has not only a description of some aspect of human behavior, but also a reason for the behavior.

VII. COPING WITH COMPLEXITY

A critical need for all studies in language understanding is effective mechanisms for coping with the complexity of the phenomenon we are trying to understand and explain. The models that are required for describing human language performance are more complicated than the comparatively simple physical phenomena in most other areas of science. Only the models in artificial intelligence and computational linguistics, and perhaps some kinds of theoretical chemistry reach the level of having theories which comprise thousands of rules (or equations) that interact in complicated ways. If the results of detailed studies of linguistic phenomena are to be disseminated and the field is to grow from the exchange of information and the continued accumulation of a body of known facts, then the facts must be capable of being communicated. We have then, at the core of the methodology of language understanding research, a critical need for some of the byproducts of our own research -- we need to develop effective formalisms for representation and for communication of our theories. The expression of a theory of language in a formal system which is incomprehensible or tedious to comprehend will contribute little to this endeavor. What is required then, as a fundamental tool for research in language understanding is a formalism for expressing theories of language (involving large numbers of elementary facts) in ways which are cognitively efficient -- i. e. which minimize the intellectual effort required to grasp and remember the functions of individual elements of the theory and the way in which they interact.

A good example of cognitive efficiency of representation occurs in the representations of transition network grammars, compared with the intermediate stages of a transformational derivation in a conventional transformational grammar. It is well known, that humans find it easier to remember lists of familiar elements which fit together in structured ways than to remember dynamically varying lists of unfamiliar things. In a transition network grammar, the stages of intermediate processing of a sentence proceed through a sequence of transitions through named states in the grammar. Each of these states has a name which has mnemonic value and corresponds to a particular milestone or landmark in the course of processing a sentence. A student of the language or a grammar designer or someone studying someone else's grammar can become familiar with each of these states as a known entity, can remember it by name, and become familiar with a variety of information associated with that state -- such as what kinds of linguistic constructions preceeded it, what constructions to expect to the right, prosodic cues which can be expected to accompany it, potential ambiguities and disambiguation strategies, etc. The corresponding intermediate stages of a transformational grammar go through a sequence of intermediate phrase markers which do not exist ahead of time, are not named, have no mnemonic value, are constructed dynamically during a parsing, and in general provide none of the above mentioned useful cognitive aids to the student of the grammar.

Similarly, the information remembered during the course of a parsing with an ATN is stored in named registers, again with mnemonic value, while the corresponding information in a transformational intermediate structure is indicated solely by positional information in the intermediate tree structure with no such mnemonic aid, with an attendant difficulty for memory, and with the added difficulty that it is possible to construct a structure accidentally which matches the input pattern of a rule that one did not intend it to activate. The chance of doing this accidentally with a mnemonically named register or condition is negligible.

Many other techniques for expressing complicated systems with cognitive efficiency are being developed by programmers in sophisticated languages such as INTERLISP, where some programmers are adopting styles of programming which make the understanding of the program by human programmers and students easier. A major technique of these programming styles from the standpoint of cognitive efficiency is the use of a hierarchy of subroutines with specified function and mnemonic names to produce program structures which match closely the human conceptual model of what the program is doing. In such systems, one can verify the successful operation of an algorithm by a method called recursion induction, which effectively says that if all of the subroutines do the right thing, then the main routine will also do the right thing. If one is sufficiently systematic and careful in his programming style, then the assurance that each level of the program does the right thing can be guaranteed by inspection and the chances of writing programs with hidden bugs or complicated programs whose function cannot be easily understood is greatly reduced.

As an example, consider a technique which I use extensively in my own programming in LISP. Suppose that I have a data object called a configuration which is represented as a list of 5 elements and the second element of the list is the state of the configuration. It is a simple matter of

programming discipline to extract the state name from such a data object with a function called CONFIG.STATE rather than the LISP function CADR, with the result that the program is almost self documenting instead of incomprehensible. It is easy in INTERLISP to define the two functions identically and to cause them to compile identically so that no cost in running time is necessitated by such programming techniques. (In my case I have a LISP function called EQUIVALENCE which takes care of all the details if I simply call (EQUIVALENCE (CONFIG.STATE CADDR)).) Recently, new features have been added to INTERLISP which further facilitate such programming conventions by providing the user with a generalized facility for record naming and field extraction.

Another example of the principle of cognitive efficiency arises in the now famous go-to controversy of the programming language theorists. One school argues that one should program in a structural discipline which makes go-to instructions unnecessary and that such a discipline should be forced on a programmer because the code he will write under such a discipline will be better. This extreme point of view is presumably in contrast to the situation in the language FORTRAN where one can handle branching only by "naming" each of the branches with (unmnemonic) numeric labels and specifying go-to instructions in terms of such labels. However, I would argue that in many other situations, with a language which permits mnemonic labels, a programmer can insert a go-to instruction for the same kinds of reasons that he creates many subroutines -- i.e., there is a significant chunk of operation which in his mind is a unit (for which he has or can coin a name) and which he would like to represent in his code in a way that will enable him to read portions of the code at a level of detail which is cognitively efficient. When go-to instructions are used in this way, they have the same value that the ability to write subroutines provides (not only efficiency of writing a given portion of code once while being able to enter it for execution from several places, but also the cognitive efficiency of being able to ignore details of how some process operates by referring to it by name or label in situations where it is the purpose or goal of a procedure or block of code which is important and not the details).

VIII. THE NEED FOR A COMPREHENSIBLE FRAMEWORK

Not only must the individual rules of a complex system be comprehensible to the system designer and the student, but also the control framework into which these rules fit must be understood. Again, there is a principle of cognitive efficiency in operation. A control framework which is simple to explain and easily remembered by the student of the system as he studies it, is far preferable to one which constantly misleads the student into thinking that something happens in one way when it actually happens differently or not at all. One cannot write rules for a system when he is not sure how it will apply the rules or when. Languages which take away from the programmer the burden of specifying the details of control structure should not also take away his ability to easily understand and forsee what will happen in response to his rules.

IX. COGNITIVE EFFICIENCY IN GRAMMARS

One of the dilemmas of the field of computational linguistics has been the difficulty of evaluating the quality of a grammar which someone has written. What is the scope of grammatical phenomena which it covers? It is one thing to say that a grammar handles questions, imperatives, comparatives, adverbs, etc. It is another thing to discover that what this means is that certain yes/no and simple wh- questions are handled, that a certain class of comparatives (the easy ones) are handled, and that only single word adverbs before or after the main verb are handled. A list of phenomena supposedly dealt with is obviously not sufficient.

A common attempt to specify the class of sentences accepted by a grammar is to list a sample set of the sentences covered. This tends to give a feeling for what the grammar can handle, but depending on the scrupulousness of the author in pointing out the things that his grammar doesn't handle (assuming he realizes what it doesn't handle) it is very easy for the reader to overgeneralize the range actually handled. When the author lists several examples of different kinds of comparatives, how does the reader decide whether all possibilities are handled or just those cases that are listed. The problem is that what one wants is a precise, compact, and comprehensible representation of exactly the class of sentences which are acceptable and how they are handled. But, notice that to the extent that such a specification is realizable, that is exactly what a grammar should be. Hence, the thing that is needed is a formalism for grammar specification which is precise, compact, and comprehensible to a human grammar designer. In short, we need a formalism for grammar specification which is cognitively efficient -- enough so that a grammarian can tell by inspection of the grammar whether a sentence is acceptable or not. While this may not be realizable to this extent, it seems to focus on the hopelessness of attempting to find some other specification of what a grammar does which will somehow be clearer than the grammar itself. Instead, it shifts the emphasis to making the grammar formalism sufficiently perspicuous that one can study it and understand it directly.

The only other method I know of at the present to obtain answers to specific questions about what a grammar does is to get your hands on the system and probe it with your theories of what it handles and what it doesn't. This has its own

disadvantages in the other direction, since it is indeed possible for a sentence to fail for a trivial reason that is a simple bug in a program and not because the grammar is incorrect or the theory is inadequate. Moreover, it is almost impossible for anyone but the designer and implementer of the system to tell whether it is a simple bug or a real conceptual difficulty and one certainly can't simply take on faith a statement of "Oh that's just a bug." However, I think that it is inevitable that natural language grammars will reach a level of complexity, no matter how perspicuous one makes the grammar, where computer aid in checking out theories and finding out what is or is not handled is an essential tool. This does not obviate the need for cognitive efficiency, however.

To make the matter more complicated, in many systems, now, the syntactic component is not separable from the semantics and pragmatics of the system so that a sentence can fail to be handled correctly not only due to incorrect syntax (i. e. the grammar does not match the reality of what people say) but also due to concepts which the system does not know or things which the system finds inappropriate to the context. For such systems, it is almost impossible to judge the capability of the individual components of the system in any objective and non idiosyncratic terms. Each system is unique in the scope of what it is trying to do and finding any equivalent grounds on which to compare two of them is difficult if not impossible. The ability to understand the formalism in which the author expresses his theory and presents it to the world is critical.

X. CONCLUSION

In conclusion, the major thrust of this paper has been to stress the complexity of scale which must be dealt with in representing theories of natural language understanding and especially in communicating them to other people. My major methodological weapon against this complexity, is to develop specification languages and notations which are cognitively efficient in the sense that they minimize the human intellectual effort necessary to understand, remember, design, and use such formalisms. We should strive for notations that can be used to publish grammars, semantic specifications, and knowledge bases in a form that one can realistically expect other people to read and understand. Simple things such as naming functions with names that will invoke the correct concept in the head of the person studying the formalism (rather than a clever name the author fancies, or the first thing he happened to name it, or the name it used to have when he used it for something else, etc.) can make an enormous difference in the cognitive efficiency of a formalism. In short, I am making a plea for making the specification language used for theory development in natural language understanding be a communication language intended and engineered for human comprehension as well as mechanical implementation. In addition, I have discussed the need to perform research in the specialized areas of language understanding within the framework of a global picture of the entire language understanding process. I have called for more care in the precise use of terms and the use where possible of accepted existing terms rather than inventing unnecessary new ones. I have also stressed the necessity that models must produce some overt behavior which can be evaluated, and have noted the desirability of finding explanatory models rather than mere descriptive models if one is really to produce an understanding of the language understanding process. I hope that the paper will serve as a useful basis for discussion.

REFERENCES

Becker, J.D. "An Information Processing Model of Intermediate-Level Cognition," Memo AI-119, Stanford Artificial Intelligence Project, Stanford University, Stanford, Calif., May, 1970.

International Joint Conference on Artificial Intelligence, London, England, September, 1971.

Woods, W.A. "Transition Network Grammars for Natural Language Analysis," Comm. ACM, Vol 13, No. 10, (October, 1970).