COGNITIVE SCIENCE 7, 95–119 (1983)

Beliefs, Points of View, and Multiple Environments*

YORICK WILKS
Cognitive Studies Centre
University of Essex

JANUSZ BIEN
Institute of Informatics
Warsaw University

The paper describes a system for dealing with nestings of belief in terms of the mechanism of *computational environment*. A method is offered for computing the beliefs of A about B (and so on) in terms of the systems existing knowledge structures about A and B separately. A proposal for *belief percolation* is put forward: percolation being a side effect of the process of the computation of nested beliefs, but one which could explain the acquisition of unsupported beliefs. It is argued that the mechanism proposed is compatible with a general *least effort* hypothesis concerning human mental functioning.

INTRODUCTION

This paper presents a model of beliefs for computer understanding of natural language and discusses its implications for speech act theory.

Although using knowledge for language understanding is an artificial intelligence (AI) tradition, the relevance of speaker's knowledge about the hearer (and vice versa) was appreciated only recently in the research of the Toronto group (Allen & Perrault, 1978; Cohen, 1978). With the exception of Bien's multiple environments approach to natural language (Bien, 1976 a, b, 1977), modelling the beliefs of the persons only mentioned in the text was completely neglected.

*The first author, during the writing of the paper, received research support from the Leverhulme Trust which he gratefully acknowledges. Correspondence and requests for reprints should be addressed to Yorick Wilks, Cognitive Studies Centre, University of Essex, England.

The authors are indebted to comments and criticisms from Dan Dennett, Bill Mann, Bob Balzer, Bob Abelson, Roger Schank, Mike Rosner, and Richard Young. The errors, of course, are all our own.

The following dialogue is perfectly natural:

USER: Frank is coming tomorrow, I think
SYSTEM: Perhaps I should leave (I)
USER: Why?
SYSTEM: Coming from you that is a warning
USER: Does Frank dislike you?
SYSTEM: I don't know, but you think he does and that is what is important now. (II)

It is clear that to follow this dialogue it is necessary to distinguish the user's beliefs about Frank's beliefs from the system's beliefs about Frank's beliefs, and from Frank's actual beliefs. Such a situation is sufficiently common to deserve special attention.

In this paper we want to tackle the issue generally and to ask the question "What is it to maintain a structure, not only of one's beliefs about the inanimate world, but about beliefs about other individuals and their beliefs?" The argument of the paper will be that there can be a very general algorithm for the construction of beliefs about beliefs about beliefs or, if you wish, models of models of models, or points of view of points of view of points of view.

Its a philosophical cliche that understanding of language is dependent upon not only the beliefs of the understander about the world, but his beliefs about the beliefs of the speaker and the ways in which those two may be different. To adopt a well-known philosophical example (Donnellan, 1971, which makes the point from a speaker's point of view), a person at a cocktail party may look across the room at a lady whose name he wants to know. He does not know her name, but knows she is a teetotaller, although he sees her holding a glass with a colorless liquid and an olive in it. Since he knows she is a teetotaller, he also knows that it is not a martini. Nonetheless, he wants to ask the person next to him who the lady is. He could ask who is the lady drinking a glass of water, which would be consistent with his own beliefs, but in fact what he says to his hearer is "Who is the lady over there drinking the martini?" He does this in order to get the reply he wants, and does it by assuming a belief which he in fact believes to be false, but which he believes to be *consistent with the beliefs of his hearer.* That is to say, he believes that his hearer knows the name of the lady in question, but does not know that she is a teetotaller and therefore assumes that she is drinking what she appears to be drinking. It is a somewhat labored story but it makes the point rather well: that we often operate with assumed beliefs which we do not in fact hold, but which we attribute to our hearers. In this paper we are going to discuss the construction of such entities as the structured beliefs of others and how we manipulate and maintain such entities.

The present paper does not describe the working of actual programs, but presents work in the course of being programmed. After we started this

work, we discovered the work of the Toronto group (Perrault, Cohen, & Allen, 1978). However, we think our proposals differ from theirs in significant respects, and at the conclusion of the paper we shall make clear what those differences are. The distinctive features of what we propose will be:

(a) The metaphorical use of the computer science notion of *multiple environments* for representing the beliefs and their interrelations: interpreting an utterance according to someone's beliefs is viewed as running or evaluating the utterance in the appropriate environment.

(b) Another crucial aspect of our approach is that a belief manipulating system, which is to be psychologically and computationally plausible, must have built into it some limitations on processing, so as to accord with the fact that deep nestings of beliefs (while well formed in some "competence" sense) are in practice incomprehensible. Consider just a small part of an example in R.D. Laing's *Knots:*

Jack thinks
he does not know
what he thinks
Jill thinks
he does not know etc. etc.

We intend our proposals to reflect this aspect of language processing (largely neglected in more logic-based approaches) that follows from real-time understanding with limited resources.

We have independently argued in the past (Bien, 1977; Wilks, 1975) for a "least-effort" view of language processing, and the present proposals are consistent with that, as we shall show.

(c) Our means for explicating this will be what we shall later call the *percolation effect,* a method in which beliefs propagate about a belief system in a way not necessarily intended by any believer or participant, but which follows as a side-effort of our principal algorithm.

The form of presentation in the paper is discussion of the design and operation of a hypothetical language understanding system capable, in principle, of performing the role of either of the participants in the dialogue quoted above.

The task we are describing is one of explicating dialogues like this, and in particular, the appropriate responses from the system (though the same methodology should apply to a modelling of the USER).

The system has produced replies at points marked (I) and (II) in the dialogue above, and the question we shall ask is why should the system say

these different things at these different times, and what structure of knowledge, inference, and beliefs about the User and Frank should be postulated in order to produce a dialogue of this type? What we shall argue is that the system is *running its knowledge about individuals in different environments at points (I) and (II),* and the difference between them will be crucial for us. In order to explain this we shall have recourse to shorthand as follows

$$\begin{matrix}\{\ \text{FRANK}\ \} \\ \text{SYSTEM}\end{matrix}$$

to represent what the bearer of the outer name believes about the bearer of the inner name, that is to say, what the system believes about Frank. Structures like this can be nested so that the following structure

$$\begin{matrix}\left\{\begin{matrix}\{\ \text{USER}\ \} \\ \text{FRANK}\end{matrix}\right\} \\ \text{SYSTEM}\end{matrix}$$

is intended to be shorthand for what the system believes about what Frank believes about the user.

We shall refer to this as a nested environment and every such structure is considered to be (trivially) inside the system, for it knows everything there is to be known about the individuals mentioned.

The first important question is, what are the structures that this shorthand represents? For the moment, the simplest form of what the system believes about Frank, i.e.;

$$\begin{matrix}\{\ \text{FRANK}\ \} \\ \text{SYSTEM}\end{matrix}$$

could simply be thought of as a less permanent version of a *frame* (Charniak, 1978; Minsky, 1975), or more suitably in the terms of Wilks (1977) as a *pseudo-text* or, if you prefer, any knowledge structure whatsoever about the individual named inside. The simplest metaphor is that of a can into which all incoming information that the bearer of the inner name is thrown. This information will have internal structure (and we shall come to that) *but the nature of the internal structure is not crucial to the argument of this paper.* The advantage of using the word pseudo-text (PT) and what it was defined to mean in that paper, is in the episodic tradition of viewing memory: that the structures of information about Frank are unsorted, unrefined (and in that sense, unframelike) items of knowledge which have not been reclassified and checked against permanent, semantic, memory. One could further this point by saying that the knowledge structure held by the sytem about Frank is in some sense only a narrative about Frank. It can be thought of as a text representation, and the earlier (Wilks, 1977) paper argued that input structures from a semantic parser of natural English could

themselves be reasonable memory structures for certain well-defined purposes.

The PT's are packed into memory schemata together with topic-specific inference rules, and the difference between pseudo-texts and inference rules may often be neglected.

Again, it is a strong assumption that the representation of the system's beliefs about entities (humans, etc.) and their beliefs is all in the same format as more structural beliefs of the system; about itself and its own functioning. We shall, therefore, propose a very general inference engine that will run over PT's on any topic, and in PT's nested to any depth, to yield an inner environment.

A further feature of this approach will be the context-dependent nature of descriptions within PT's (cf. Bobrow & Norman, 1975; Norman & Bobrow, 1979), and their associated pointers.

The context dependency of descriptions originally meant that they were never more precise than was needed in the context of their creation, but we now understand the feature in a broader sense: a given description in various contexts may refer to different times. For example, "the murderer" in the environment of John's beliefs may refer to Jones, but, in the environment of May's beliefs, to Smith, In other words, the context-dependent descriptions supply us with some power of intensional logic, which is necessary for an adequate knowledge representation, although in the simplified example we use for illustrative purposes below, we shall not make use of this feature.

CONSTRUCTING ENVIRONMENTS

The essence of this paper is to evaluate and compare two perspectives or environments (equals nested PTs) and they will be the ones which are created by the system at points II and II in the dialogue above. "Evaluate" as it is used here is intended to have a standard computer science meaning—one we could put more adventurously as *running structural descriptions in given environments* (Bien, 1977). What this will mean in concrete terms, is to draw plausible pragmatic inferences, and in that sense our view of understanding is to be identified with the drawing of such pragmatic inferences in context. This is the standard tradition of the AI approach to natural language of the last 10 years.

In particular, at (I) in the dialogue, the system is evaluating the user's initial remark 'Frank is coming tomorrow, I think' in the following nested environment

$$\left\{ \left\{ \left\{ \begin{array}{l} \text{SYSTEM} \\ \text{FRANK} \end{array} \right\} \right\} \right\}$$
$$\text{USER}$$
$$\text{SYSTEM}$$

whearas, at Point (II) in the conversation, the system has evaluated just Frank's view of himself, that is to say, he has run the user's first sentence in the simpler environment

$$\left\{ \left\{ \begin{array}{l} \text{SYSTEM} \\ \text{FRANK} \end{array} \right\} \right\}$$
$$\text{SYSTEM}$$

where he discovers that he has no such information on what Frank thinks of him.

If we suppose some parsing of the input sentences into a semantic representation, the first question of principle is that of strategies for setting up environments. We shall distinguish the *presentation strategy* and the *insertional strategy*. The question for the presentation strategy is this: given any incoming information about an individual, how deep a level of nested environment should the system construct? A minimal strategy will be appropriate when, for instance, listening to a mathematics lecture, where one is not normally evaluating the input in terms of the presence of the speaker: one is not asking oneself "Why is this algebra lecturer telling me this?" and evaluating his motives and reasons for doing it. We shall call that a *minimal strategy* that would have a very shallow environment with no level corresponding to the speaker.

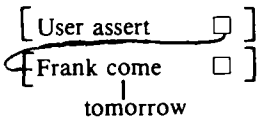
What we shall call the *standard presentation strategy* for information is the one adopted in the nested environments above, where they are nested so as to include a level corresponding to the speaker (the user in this case) and then the individuals mentioned. At (I) the system evaluates the initial sentence while taking account of the speaker's motives, but at Point (II) he does not take account of the speaker and the speaker/user does not occur in the nesting.

This standard strategy allows a hearer either to disbelieve a speaker or to cooperate with him, as he chooses (cf. Taylor & Whitehill 1981).

In addition to this, we can imagine super-strategies (see further below) which are reflexive nestings of speakers and others, as well as the system itself. In these the system constructs even deeper environments corresponding to what it believes about what somebody else believes about what it itself believes, etc. We shall call anything deeper than a single nesting using the speaker a reflexive strategy, and we shall not consider it here.

The second important question concerns what we have called *insertion strategies*. When the system reads something as at (I), said by the user about Frank, the question arises as to where in the system this should be stored. In

the case of the user's sentence "Frank is coming tomorrow, I think", should the semantic representation of that sentence be stored simply in the pseudo-text for the user, the person who said it, or in one for Frank, the person spoken about, or both? For reasons that will become clear later, we shall adopt a "scatter-gun" strategy, in that the information will be stored initially in all the possible places, i.e., both the PT for user and for Frank. This is an assumption that may need revision in the light of later experiment, but the system we are describing is heavily orientated toward redundant storage of information. For the form of the information to be stored, we shall simply assume some simple standard semantic parsing (Wilks, 1975) so that the user's statement "Frank is coming tomorrow" will parse as the following structure of simple templates



where each wor-like item above is itself a pointer to a complex semantic representation, and the parsed structure above, for a simple dependency of two clauses, indicates the agents (first slot) and actions involved (second slot), the objects being dummies (□). This whole structure will be added into *both* { USER } and { FRANK } . Since all PT's are in SYSTEM, it is not necessary to add this to lowest level PT's.

Notice here that the PT's are general items and will not be stored only for individual human beings, but also for groups of humans, objects, substances, classes, my car, a jury, a professor, a salesman, sulphur, and Germany. In Wilks (1977), their hierarchical relations and inheritance relations were discussed, and here we may assume that these are standard. In this paper we concentrate only on PT's for agents (we shall explain why later), and a consequence of this is that when we consider nested environments, PT's for agents will be the only ones that can be outer environments in nesting diagrams. That is, we can consider computing, for example, Jim's view of the oil crisis, but we cannot consider the oil crisis' view of Jim. We can do this for groups as well: we are able to consider Germany's view of the oil crisis, although never the oil crisis's view of Germany. In the examples in this paper, we shall confine ourselves to the names of individuals rather than groups or states or classes of individuals as names on the "outside" of PT's.

BELIEFS ABOUT AND BELIEFS OF

A final preliminary distinction we must make is between someone's beliefs *about* someone and his beliefs ~~ab~~out the beliefs *of* that individual. To put it

simply, we can have beliefs about Smith—that he is male, 45, etc. We can also have beliefs about his beliefs: that Smith believes that Vitamin C cures colds. On one general view of belief these are all properties of Smith from the believer's point of view, but they are, of course, importantly different sorts of property.

Earlier, we discursively introduced the knowledge structures (PTs) about animate and inanimate entities. Before we get to the heart of our paper, which is an algorithm for constructing an *environment,* or point of view that one of the entities represented by a PT can have of another, we must discuss in a little more detail that structure of the PTs. In Wilks (1977) we introduced a PT as a narrative-like structure, within which was collected, in a semantic representation, the information the system had about some entity or generic class of entities. It was considered separate from a semantic definition, as well as from a frame, which was a permanent memory structure (largely of episodic information). A PT was intended to be an intermediate memory structure, some of whose contents would undoubtedly be transferred to permanent memory. In Wilks (1977) we expressed information about the generic concept of CAR, separate from the definition of a car as a people-moving device (a nonessential separation, and KRL and Charniak (1978) would have chosen differently at this point), and containing material appropriate for a permanent frame (CARs have fluid injected into them, so that. . . .etc.), as well as episodic material that would not be so transferred (MY CAR is purple). The simple illustrative PT of that paper used the symbol * for "car" in its templates because it was not a pointer to another definitional semantic formula, PT, or associated frame (as WHEEL would be), but to the PT itself in which it occurred. It was thus a special pointed carrying, as it were, a warning against vicious self-reference.

In the present case, where individuals are concerned that can themselves have beliefs, we must amend the notation. Suppose we ask where the system keeps its knowledge about Frank. He, like the car, will have a semantic definition (human, male, etc.), as well as beliefs the system has about him (Frank is an alcoholic), as well as beliefs the system has that it explicitly believes to be Frank's beliefs (Frank believes he is a robot, and that he is merely a social drinker). Thus Frank's beliefs, as known to the system, can superficially contradict both definitional and accidental information about himself, if we may use an old fashioned distinction here (again, nothing crucial depends on it).

One also has a pretty firm intuition that the structure of Frank's own beliefs (as believed to be his by the system, of course) are rather different from beliefs *about* him. And yet if, for administrative and computational convenience, we want to keep the system's view of Frank in one place, these should all in some sense be in the box marked "Frank" (different though they are). There is an additional complication, which will be very important

when we come to a proposed algorithm, that many of the system's beliefs *about* Frank also, as a matter of fact, correspond to his own beliefs, even though we have no direct evidence of the correspondence. That is to say, the system believes Frank to be a human, even though, oddly enough, he believes himself to be a robot and that is known to the system. However, the system believes Frank to have two hands and so does Frank, even though the system has never heard him say so. All this will fall under the general rule we shall use later, that X's view of Y can be assumed to be one's own view of Y, *except where one has explicit evidence to the contrary;* and this rule also covers X's view of Y!

In the sample, over-simplified PT below we shall continue to indicate mention of Frank, in the PT for Frank (i.e., the system's beliefs about all aspects of Frank), by *. This is again (as in Wilks, 1977) a pointer warning against vicious self-reference. We shall not include the semantic definition in the PT (as Charniak would), and the * pointer (after Kleene's star) may be considered as pointing to the semantic formula definition, which in turn points to the PT. This is just as the "Earth" in the sample below is also a pointer to the corresponding semantic formula definition and PT, though with no risk of self-reference in this case. The belief "Earth is flat" is inserted precisely because it is an odd, and so reportable, belief. "Earth is round" would not be so inserted for a PT for a round-earther simply because it could be inherited from the lattice of common knowledge from the PT for EDUCATED HUMAN.

Thus in the sample PT below, the semantic formula definition is not present but assumed, and the horizontal line simply divides the beliefs above it (which are explicitly believed by the system to be Frank's, and so can be thought of as prefaced by an implicit * BELIEVES....) from those below which are beliefs about Frank, that may or may not be his. In some clear sense those above the line are a sort of "inner Frank" and the different function they have when we construct a "push-down" algorithm to create environments, will become clear.

So, the following might be trivial content for a PT $\left\{ \begin{matrix} \text{FRANK} \\ \text{SYSTEM} \end{matrix} \right\}$:

$$\left\{ \begin{array}{l} \text{FRANK} \\ \text{Earth is flat} \\ \text{System likes*} \\ \text{User dislikes*} \\ \hline \text{User dislikes*} \end{array} \right\}$$

SYSTEM

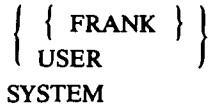
The line across the PT separates beliefs believed to be *of* Frank (above the line) from those believed *about* Frank (below it). We shall have to exercise great care with the line because, in practice, some below-line beliefs will be

believed by their object: if I say of Frank that he hates dogs, I may put that below the line, but a simple pragmatic inference rule would lift it above.

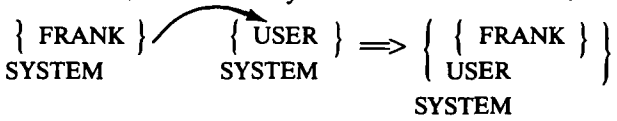
As a shorthand, above-the-line beliefs only will be *underlined.*

PUSHING DOWN ENVIRONMENTS

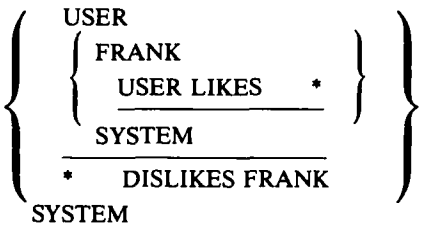
We are now approaching the heart of the paper. Pushing down one of the PTs *inside another* means resetting values in the PT being pushed down. The transitory object achieved by this method of environment we shall interpret as being the other PT *holder's* view of the inner PT object. Suppose we want to construct the system's view of the user's view of Frank



We shall assume this is done in two stages as follows: first by constructing, or having available, the system's view of Frank; second by constructing, or having available, the system's view of the user, and then pushing the former down into the latter to achieve the system's view of the user's view of Frank:



Suppose the content of the outer PT, the system's view of the user, contains the proposition "User dislikes Frank", and the inner PT, the system's view of Frank, contains the proposition "User likes Frank", believed by Frank (and hence underlined in our simple notation)

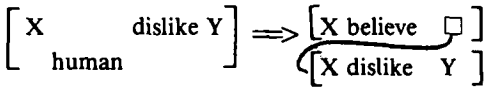


One of our major assumptions in this paper is that the system does not *preserve* complex constructions of environments; the complex points of view. It maintains structures only at the bottom level: in terms of our earlier crude metaphore, it maintains simply a row of PTs about individuals (and other entities) and not environments, i.e., not push-downs of points of view. If the system, by pushing one of these down into the other as we postulated, wishes to construct the inner environment which, in this case, is the system's

the outer belief on the inner belief. We shall consider outer beliefs migrating or *percolating* into the inner environment and then examine their mutual interaction.

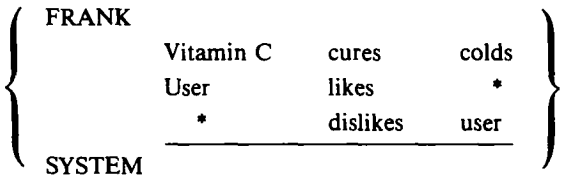
However, it should be clear that only "upper half" templates of a PT can migrate into the inner environment and override what it contains. If I want to consider what User thinks of Frank, I will let my beliefs about the User's view of Frank override my views of Frank, but not, in general, my views of the User.

However, the matter is not quite so simple for certain attitude verbs like dislike. If I say X dislikes Y, I imply X believes he dislikes Y except in exceptional circumstances, or, in a simple rule: for a class of "conscious attitude" verbs, we have:

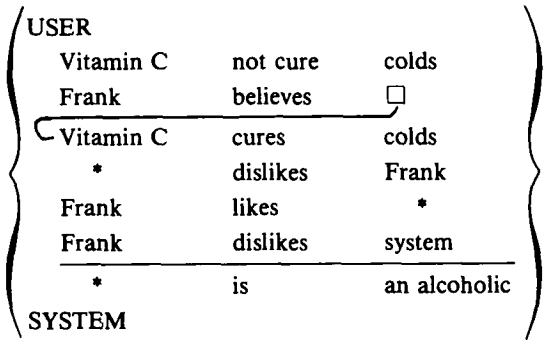


Again, if I believe Smith is 6 feet tall, I imply he believes he is 6 feet tall, as much as I assume he accepts his own "semantic definition" (male, human, etc.). We shall return to the problem of self-knowledge later.

Let us now consider in more detail the push-down of the system's view of Frank into the system's view of the User, and assume the following propositions as constituting a slightly more complex PT for the system's view of Frank.

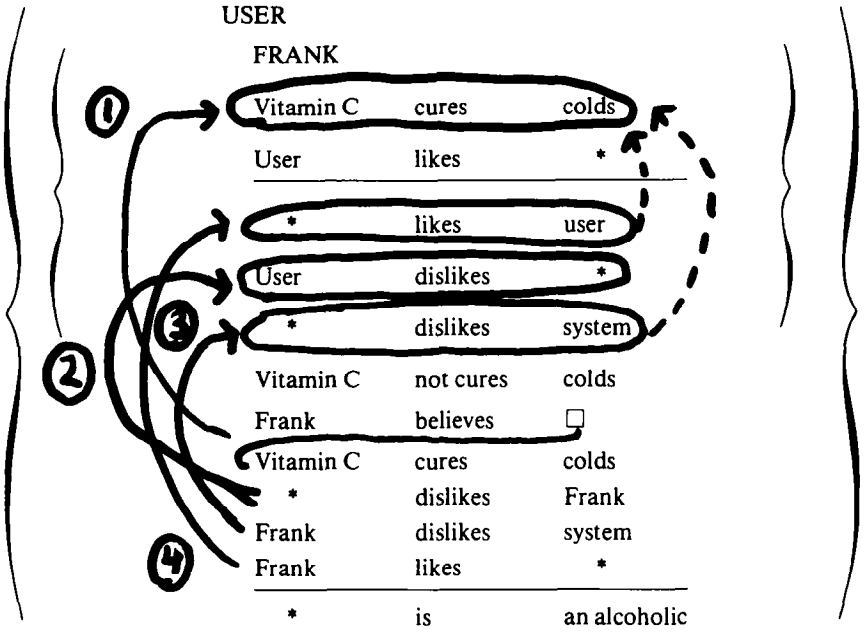


Now consider this in relation to a more complex view of the system's view of the user:



In this example we see the more complex two-clause proposition: the one that expresses the system's belief that the user believes Frank believes that Vitamin C cures colds. If we now consider the first of these to be pushed

down into the second, we achieve a complex item like the following example in which the arrows show the interaction as the relevant outer propositions *percolate* into the inner environment. This is fundamentally simple, and we now want to discuss one-by-one the interactions that are obtained by this method. We have numbered the linkages 1, 2, 3, and 4 to show four separate percolations "inwards" (templates shown ringed).



The two dotted lines show those templates that mount to the upper half of { FRANK } via the inference rule. Other percolations remain in the lower half, except the one marked 1 whose content requires the upper-half beliefs.

What we have to consider is those propositions in the outer PT being drawn into the inner pseudo-text by a relevance criterion, given that they are above the line in the outer PT, for only *explicit beliefs* of the user can override anything inside the (bottom half of) the inner PT. Our diagrammatic distinction of the upper and lower halves of the PT is thus a notation for limiting inference. We shall discuss the criterion that causes it to be considered relevant further below but, for the moment, we take an over-simple criterion of explicit mention in the outer PT (upper half) of the "holder" of the inner PT.

1. We have the proposition "Frank believes that Vitamin C cures colds." This enters the inner PT and in appropriate-reduced form, "Vitamin C cures colds" is just another copy of the system's belief about Frank.

2. In the inner PT

$$\begin{matrix}\{\ \text{FRANK}\ \}\\ \text{SYSTEM}\end{matrix}$$

we have the system's belief that Frank believes that the user likes him, and in the outer PT we have the relevant belief of the system of the user that the user dislikes Frank. The latter comes in to the inner PT as "User dislikes Frank," but into its bottom half, not immediate as a belief of Frank. There is, of course, no contradiction between these, because the inference rule given does not generate any contradiction.
3. The belief in the outer PT "Frank dislikes system" is drawn into the inner environment and meets no contradiction of any kind so it simply remains there. We shall have cause to refer back to this later as a crucial percolation.
4. We have in the outer PT the belief held by the user that Frank likes him. This is drawn in and, given the inference rule above, contradicts the inner belief of Frank that he, Frank, dislikes the user. The inner belief that he, Frank, dislikes the user is explicitly contradicted and overwritten by the outer belief Frank likes user. The *s change reference, since they refer to the local subject of the PT, the incoming belief taking the form "* likes User".

The result of these four operations in the inner environment is now the construction

$$\begin{matrix}\left\{\begin{matrix}\{\ \text{FRANK}\ \}\\ \text{USER}\end{matrix}\right\}\\ \text{SYSTEM}\end{matrix}$$

that is to say, the system's view of the user's view of Frank. So we now have an inner result as shown above in the inner environment.

The reader will recall the two heuristics used in order to obtain this intermediate result. First, the *relevance heuristic:* the basis for considering outer PT propositions (from the "upper" explicit belief part) as being possible migrants into the inner environment. The criterion has been explicit mention, of Frank in this case, in the outer propositions. The reader should not assume that the beliefs listed in the two example PTs were chosen as relevant to the example: the original example dialogue referred to Frank hating the user, which was in a PT, but there was nothing relevant to the example about his views on Vitamin C and its efficacy. Those were drawn into the inner environment simply by the general relevance mechanism. There will, of course, be considerable practical problems caused by conjunction, disjunction, etc. given a richer PT structure. Moreover, the present criterion of relevance (explicit mention) is far too naive: in fact, there will not nor-

mally be relevance only by name, but also by descriptions. So, for example, Frank may not appear in the outer propositions as his name "Frank", but under some description such as "John's father," in which case much more complex inference procedures will be required to establish the relevance of the corresponding outer proposition. This is a general problem for all artificial intelligence systems of this type and no peculiar difficulties arise here. Second, *contradiction:* inner beliefs survive unless contradicted by outer ones from the upper, explicit, half. Any inner beliefs which is not contradicted survives.

The reader may ask at this point what is the function of the heuristics and construction we have produced, and upon what general intuition do they rest. The assumption is that of basic common sense: that one's view of another person's view of a third person (and so on) is simply that it is the same as one's own view *unless one has explicit reason to believe otherwise.* This sounds extremely simple but, as we have seen, there are considerable complications in working out even this very common sense principle.

At this point let us recall why we have performed this environment construction. The initial example of dialogue required the first environment (at the point marked (I)).

$$\left\{ \begin{array}{l} \left\{ \begin{array}{l} \{ \text{SYSTEM} \} \\ \text{FRANK} \end{array} \right\} \\ \text{USER} \end{array} \right\}$$
$$\text{SYSTEM}$$

This is obtained in two such push-down moves. First,

$$\left\{ \begin{array}{l} \{ \text{FRANK} \} \\ \text{USER} \end{array} \right\}$$
$$\text{SYSTEM}$$

which we have just done, and then the full nesting by pushing down

$$\{ \text{SYSTEM} \}$$
$$\text{SYSTEM}$$

into that. After applying these two moves, we shall achieve the inner environment required at point (I) in the example. Let us now consider two effects of the push-down environment achieved. First, within

$$\left\{ \begin{array}{l} \left\{ \begin{array}{l} \{ \text{SYSTEM} \} \\ \text{FRANK} \end{array} \right\} \\ \text{USER} \end{array} \right\}$$
$$\text{SYSTEM}$$

we now have the percolated proposition "Frank dislikes system", and we also have from our parsed input "User asserts that Frank comes tomor-

row." Notice that this belief would not have percolated into the alternative environment

$$\left\{ \begin{array}{l} \{ \text{SYSTEM} \} \\ \text{FRANK} \end{array} \right\}$$
SYSTEM

constructed at point (II) in the example, because the belief originated in

$$\{ \text{USER} \}$$
SYSTEM

which is not involved here.

Therefore, if we now have appropriate inference rules we can generate:

"Perhaps I (system) should leave" in the innermost environment, if we suppose[1] some common sense inference rule of the following type:

$$\begin{array}{ll} & [X_{\text{HUMAN}} \quad \text{DISLIKE} \quad *] \\ & [Y_{\text{HUMAN}} \quad \text{ASSERT} \quad \square] \\ & [X_{\text{HUMAN}} \quad \text{APPROACH} \quad *] \\ \Longrightarrow & [\ * \quad \text{LEAVE} \quad X] \end{array}$$

We shall not defend specific inference rules here, but only assume that they have a function in general semantic parsing systems of the present type; similar rules exist in many other systems. If we now match this rule onto the contents of the inner environment we have, we shall see the first line matches with the proposition obtained through push-down, "Frank dislikes the system". The rest of the lefthand side matches the parsed input: a human Y, which is the user, asserts the human X, which is Frank, approaches the object, which is the system, the * since we are in a system PT (the innermost). From these we infer, by rule, the output "*leaves," that is to say, the system leaves. Thus we have the motivation of the system's reply at (I) in the original dialogue. This inference rule could appear as a partial plan in a more fully-developed speech act system.

In any actual operation it would probably help to have a rule, such as the one above, stored as part of a class of rules which we might label WARN. It is here that the belief manipulations of this paper bear upon the general topic of speech acts. If it were procedurally useful to have a class of such rules stored under a primitive lable like WARN, then it would give justification for the use of speech act terminology as part of belief systems of this type. The problem is that this would be a very weak defense of

[1]See Section 6.

speech acts because, as with all primitives (like WARN in this case), it can be argued (Charniak, 1975) that the taxonomy of items one proposes could be maintained without primitive labels. Per that view, one could have a classification of inference rules available, but would not be helped further by special labels for the classes (such as warn, threat, promise, assert, etc.) in order to access that rule class.

PERCOLATION

A second result of the push-down we have arrived at is that the proposition "Frank dislikes the system" can now be retained in the inner PT *if we adopt what we shall call the "percolation heuristic"*. This is as follows: when a proposition has appeared in an inner environment, and is not contradicted, it remains there in the standard copy of that inner PT *when that copy is subsequently re-established at bottom level.* We shall say the proposition "Frank dislikes the system" has *percolated* from the "outer" PT

$$\begin{matrix} \{\ \text{USER}\ \} \\ \text{SYSTEM} \end{matrix}$$

into the "inner" PT

$$\begin{matrix} \{\ \text{FRANK}\ \} \\ \text{SYSTEM} \end{matrix}$$

After the example dialogue has been dealt with, the inner PT is, as it were, pulled out and remains the system's view of Frank. The percolation heuristic which asserts that "Frank dislikes the system" *stays in that copy*[2] *for the future.* By iterative percolation, it will also have gone further into the innermost environment

$$\begin{matrix} \{\ \text{SYSTEM}\ \} \\ \text{SYSTEM} \end{matrix}$$

as "Frank dislikes the system". We shall argue that on a least-effort principle of belief manipulations, this percolation heuristic is justified. But first let us consider an immediate potential counter-example. Supposing the user had said "Frank thinks you're crazy, but I don't". We might imagine a contradiction achieved by percolation into the innermost environment (which is the system's view of the user's view of Frank's view of the system), where we might have mutually contradictory percolations: both "The system is crazy", and "The system is not crazy." In fact, if we follow this in detail, we find this does not happen: the worst that can happen is a contradiction in

[2]This allows the possibility of multiple copies of a PT for Frank and hence the problem of which is most *salient* in later retrieval (see Bien, 1980).

the system's view *of the user's beliefs* after percolation. Moreover, none of us find anything disturbing about the idea that others (rather than ourselves) have contradictory beliefs. Nevertheless, although no contradiction follows here, we might well want to have certain surface key words inhibit multi step percolations, in this case "but".

The principle suggested here is that percolations remain for all future use of a given PT, not just in relation to the PT from which they percolated. Another result of percolation will be that

$$\left\{ \begin{array}{c} \text{FRANK} \\ \text{SYSTEM} \end{array} \right\}$$

is no longer the *general* representation of Frank, for beliefs about him may percolate anywhere in the system's PTs.

One argument for the percolation heuristic is based on the assumption that pushing down PTs inside each other to create environments requires considerable computational, or psychological, effort; with greater effort required for greater nesting depths. Allowing beliefs to percolate about the system in the suggested way, avoids having to recompute the same environment by repeating a push-down should another dialogue by encountered that required the same environment. Before running a sentence representation in an environment, we can check the push-down nesting required and examine a flag in the innermost PT to see whether that inner environment had been constructed recently or not. The notion of recency would have to be firmly defined but, if the sentence representation had been run recently on any such definition, we would assume that percolations and settlings of consistencies had been done and would not need to be repeated.

An important point here is that the push-downs are not in any sense topic-guided: that is to say, if the text required us to calculate Reagan's view of Begin's view *of the oil problem,* then we could not be sure that, if we were to assemble the same order of nesting of outer PTs for another topic, say Reagan's view of Begin's view of Saudi Arabia, then the register of previous push-downs would mean that all possible consistencies and percolations had already been achieved in the innermost environment (since they would have been directed by relevance to oil). However, our overall least-effort principle of comprehension requires that we do not repeat the *outer part* of that push-down, which would have been constructed already.

It must be remembered that the push-down metaphor is merely a metaphor. The actual computation involved in computing the inner environments is a cross product of PT contents. However, it might turn out experimentally that the assumption here about percolation is not suitable for all topics. We might, on the basis of experiment, wish to restrict percolation itself to certain topics or psychological modes, in particular those of *attitudes.* What is being suggested here, in psychological terms, under the

metaphor of percolation, is that it is essentially a side-effect that transfers beliefs for which one does not have explicit evidence. There is a phenomenon called the sleeper effect (Gruder, et al., 1978) which is well attested, and yields experimental evidence about how people come to hold beliefs for which they have no direct evidence of any kind. We take this as indirect evidence that something along the lines we propose for percolation could in fact be experimentally tested.

PERCOLATION AND THE ABOUT/OF BELIEF DIVISION

It will be clear from the earlier detailed discussion of the sample push-down, that it is normally only the upper-half beliefs of the lower PT that migrate into the lower half of the PT being pushed down. The diagrammatic upper and lower half metaphor expresses this conveniently, but the motivation is also clear: it is, in general, the system's beliefs about the lower PT-holder's beliefs that will modify the beliefs stored in the (lower half of) the upper, or pushing-down, PT. As we saw, beliefs in the lower PT that explicitly refer to the beliefs in other PTs (e.g., USER BELIEVES FRANK BELIEVES VITAMIN C CURES COLDS) will naturally migrate to the upper or inner part of a pushing-down PT.

The upper and lower half of a PT metaphor has, of course, no application to a PT for an inanimate entity or entities (except perhaps computers) since they do not have the inner beliefs. So we can think of a PT for *coal* as having all its content below some notional line, and nothing above. Thus, nothing can be pushed down into such a PT, and when it is pushed down into some other (animate) PT (so as to construct, for example, someone's beliefs about coal) all migrations will be into its lower half. There will be the standard status problems about whether PTs for such entities as *France* and *The Auto Workers' Union* can have beliefs.

One further clarification is needed of the meaning of the percolation heuristic proper, as opposed to beliefs migrating into an inner PT on some relevance criterion. Percolations are those beliefs that migrate into a lower half of an upper PT *and are not contradicted*[3]. It is those, we suggest, that may remain in the resulting permanent copy of the upper PT.

[3]Where contradiction also covers those contradictions reached via inference rules, as in

John is married to Mary

and

Fred says John is married to Rita

inside $\left\{ \begin{array}{l} \{ \text{JOHN} \} \\ \text{FRED} \\ \text{SYSTEM} \end{array} \right\}$

the latter sentence overrides and contradicts the former, only given some rule such as

X IS MARRIED TO Y $\Longrightarrow$ X IS NOT MARRIED TO Z ($\neq$ Y)

The key example in the discussion was FRANK DISLIKES THE SYSTEM. Now, we have suggested that on withdrawal of the upper PT from the push-down, such a belief might remain, with its source believer stripped away, as it were, and so be a belief of the system that it had merely inherited as a side-effect. It will be clear that such persistence could not apply to beliefs that had migrated in and overriden inner beliefs: as FRANK LIKES USER migrated into $\left\{ \begin{array}{c} \text{FRANK} \\ \text{USER} \end{array} \right\}$ and contradicted/overrode FRANK DISLIKES USER. If that were to remain in a permanent copy, any push-down might cause the sytem randomly to reverse its beliefs: e.g., having calculated Begin's view of Reagan and finding it the reverse of its own, it would not be likely to reverse its own beliefs on Reagan just from having constructed that particular environment!

Two other important topics have been left in an incomplete state, and will require further discussion in another paper: first, what classes of beliefs can be inferred pragmatically to be those of their subject and, second and closely related, are "self-pushdowns" significantly different from the general case, i.e.,

$$\left\{ \begin{array}{c} \text{FRANK} \\ \text{FRANK} \end{array} \right\}$$

from

$$\left\{ \begin{array}{c} \text{FRANK} \\ \text{USER} \end{array} \right\}$$

We noted earlier that there is a class of attitude beliefs that we would expect to migrate from the lower to upper halves of a PT, e.g.,

X DISLIKES Y ⟹ X BELIEVES (X DISLIKES Y)

Again, one would expect the same inference to hold for:

(a) all parts of a semantic definition,[4] unless explicitly overriden;

X IS HUMAN ⟹ X BELIEVES (X IS HUMAN)

(b) all parts inherited from a "lattice of common knowledge PTs";

X IS AN ARCHITECT ⟹ X BELIEVES (X CAN-PLAN BUILDINGS)

(c) what Clark and Carlson (in press) have, following Schiffer, called *mutual knowledge;* for example, when X and Y observe a candle, an infinite number of propositions such as

X KNOWS Y KNOWS THERE IS A CANDLE

can be inferred by a simple rule of construction (in which the superficial differences between *know* and *believe* are not significant). There has been

[4]In Wilks (1977), semantic definitions were started separately from PTs, rather than together as in Charniak (1978), but nothing here depends on that.

much misunderstanding about the degree to which Clark and Marshall (1981) thought that such complex nested entities were really (rather than trivially) constructed by participants. However, it is clear that in certain complex situations, such as detective stories, independent evidence can be offered for a number of levels of such nestings in situations more complex than mere copresence with an object (such as a candle). These situations do not collapse trivially (or, conversely, are not trivially inferrable by a recursive rule) and are more like the processing of (non-collapsing) center embeddings:

The dog the cat the man saw bit died.

In those cases, deep nestings are very hard to compute and handle for subjects, and this is additional evidence supporting our "least-effort" view of environments. One could view the present paper as a beginning for a procedural account of limitations on (non-trivial) "mutual knowledge" embeddings.

(d) A natural additional inference would be;

$$\text{X ASSERT P} \Longrightarrow \text{X BELIEVE P}$$

unless there was any indication of, or reason to suspect, lying by X.

Indeed, application of this rule makes the basic inference to the systems dialogue reply far clearer because

$$\begin{array}{c} (\text{USER ASSERT} \quad \square) \\ (\text{FRANK COMES} \quad \square) \\ \text{TOMORROW} \end{array}$$

is replaced by a belief form, which enters the inner environment, and the response is then generated by a simplified, and more plausible, inference rule

$$\Longrightarrow \begin{array}{l} [(\text{HUM X}) \quad \text{DISLIKE} \quad (\text{HUM Y})] \\ [(\text{HUM X}) \quad \text{APPROACH} \quad (\text{HUM Y})] \\ [\text{Y} \quad (\text{MOVE AWAY}) \quad \square] \end{array}$$

SELF-EMBEDDING

A topic that has been touched on, but not confronted, is that of an individual's view of himself, and the ways in which this does not conform to our general heuristic for the computation of points of view: someone else's view of X is my view *except where I believe that not to be the case.*

No problem arises with the system's self-model: the PT

$$\begin{array}{c} \{ \text{SYSTEM} \} \\ \text{SYSTEM} \end{array}$$

has all its content above the line (to continue the demarcation line metaphor). There are no beliefs the system has about the system that are not its own beliefs.

More interesting cases arise when the system wishes to compute Frank's view of himself or Frank's view of the system: i.e.,

$$\left\{ \begin{array}{l} \{ \text{FRANK} \} \\ \text{FRANK} \end{array} \right\}$$
SYSTEM

or

$$\left\{ \begin{array}{l} \{ \text{SYSTEM} \} \\ \text{FRANK} \end{array} \right\}$$
SYSTEM

It would be reasonable to assume that Frank's view of himself is, *in general,* the same as my view of him except where I have evidence to the contrary. Our main heuristic would create an environment in which Frank believes his address to be what I believe it to be; his number of eyes to be what I believe it to be; but his number of teeth to be the value of what he believes it to be (and not the empty slot that I have); and so on. In other words, Frank may well have beliefs, concrete and abstract, that I know nothing about, but given the limitations on my beliefs, my best construction is still to believe that his beliefs are as mine (except for the listed exceptions that I am aware of, and the special treatment of empty slots). So, we might say, for the first of these situations behaviorism is a safe intellectual policy.

What about the second case: the computation of Frank's view of the system? Here the general heuristic must surely break down, because the system cannot assume that Frank has access to all the beliefs about itself that it has. The situation is the inverse of the earlier one as regards the evaluation of slot fillers, for if we applied the general heuristic to a system that believed itself to be human and believed a particular figure for its number of teeth, we would construct an inner environment in which Frank would have the system's own beliefs about its number of teeth, which is not at all plausible. One simply knows oneself better than others do, and that fact has concrete expression for oneself (although that implies nothing about behaviorism and the degree to which others could, *in principle,* know as much about one as one does oneself). Though in the case of one's beliefs about others, one believes they know more about themselves than one does oneself, but nothing follows from that (if one knew what such things were they would be known to one, ex hypothesi).

However, it ought to be possible for the system to ask and answer the question: "What is Frank's view of me?", if only because this is a common and troublesome question in everyday life. If we apply the heuristic to compute:

$$\left\{ \begin{array}{l} \{ \text{SYSTEM} \} \\ \text{FRANK} \end{array} \right\}$$
SYSTEM

we would get Frank's explicit beliefs about the system (as believed by the system), and the upper half of

$$\{ \text{FRANK} \}$$
SYSTEM

migrating into the lower half of

$$\{ \text{SYSTEM} \}$$
SYSTEM

which was (see above) previously empty, since all the system's beliefs about itself are necessarily its beliefs. We must assume that Frank's beliefs about the system also include some of the list (a)–(d) above, semantic definitional knowledge, copresence information, as well as general information from the higher level PT $\{ \text{HUMAN} \}$, or of course $\{ \text{COMPUTER} \}$, if the system has been found out for what it is. However, this does not add up to much, if the inner environment is to be limited (as it seems it must be in this special case) to the *lower half* of the inner

$$\{ \text{SYSTEM} \}$$
SYSTEM

so as to avoid the system assuming Frank knows all the things about it that it does.

More thought is required on this issue, and a solution may be found (in the sense of a psychologically plausible solution) via a special PT for the system's view of its *public self* (i.e., what it believes to be the public's view of it, including what it is publically believed to believe about itself). But it would seem a pity to introduce a special entity here; one that ought to be constructible from the entities already available in the system, lest such entities have to be produced in some form for all other entities in the system. Such entities would not fit with the general assumptions of this paper because they would be essentially stored push-downs instead of everything being stored at the lowest level, as we have assumed, for the entity proposed would be:

$$\left\{ \begin{array}{l} \{ \text{SYSTEM} \} \\ \text{AVERAGE MAN} \end{array} \right\}$$
SYSTEM

A way of avoiding this would be for there to be beliefs in the lower half of

$$\{ \text{SYSTEM} \}$$
SYSTEM

and for them to be the system's beliefs about the average man's view of the system's self. If there were present, then the general heurisic would run properly.

It needs to be emphasised that all operations in the system are, in some sense, self-embedding since all the PTs are the system's PTs and are *trivially* indexed from

$$\begin{matrix}\{\text{SYSTEM}\}\\ \text{SYSTEM}\end{matrix},$$

i.e., only a minute fraction of the system's real beliefs are actually in the upper part of

$$\begin{matrix}\{\text{SYSTEM}\}\\ \text{SYSTEM}\end{matrix}.$$

But the system should be able to distinguish between what it believes about elephants, and what the average man believes about them. After all, the system might be an expert on elephants and any simple minded application of the general heuristic would again be wrong. The plausible solution here is again a pointer from the upper half of the PT $\{$ AVERAGE WESTERN MAN $\}$ to the system's own beliefs as default, but with "expertise areas" segregated in those of the system's PTs about which it is an expert (corresponding perhaps, even in the semantic definitions, to Putnam's "division of linguistic labor").

Some, on reading this, will want to argue for reverse or outward percolations, but we would claim that on detailed inspection such cases all turn out to be (inward) percolations falling under the general heuristic rule. The most obvious case would be the inverse of X believes p$\rightarrow$p, namely, p (in one of the system's PTs)$\rightarrow$Average man believes p. However, this is accounted for (without any "reverse percolation") by this default arrangement of pointers.

CONCLUSION

Enormous gaps in what has been described here will immediately be evident to the reader. First, we have said little about the nature of the organization of the inference rules and their relationship to plans. This is in part deliberate, for there are many systems for which plans are central, and planning is a relatively well understood sub-topic in AI. The specific claims of this paper do not bear on that issue. Second, we have noted that a system like this could not be serious until augmented by intensional logic notions, such as being able to show the equivalence of *Dolores* to *John's mother,* for example, or *Frank* to *Jack's father.* Without this any kind of relevance heuristic for deciding what, in an outer PT, should be allowed to percolate

into an inner, would be inadequate. Third, we have said nothing about the relationship of the intermediate (and partly episodic) PTs to permanent memory frame, but again this is a subject of study in many other systems and no specific problems concerning that issue arise here.

Finally, something should be said about the differences between this work and work on speech acts and plans done at Toronto by Perrault, Allen and Cohen (see Allen & Perrault, 1978; Cohen, 1978). One is their emphasis on plans, which are not of central concern to us. A second, and fundamental, difference is that in the Toronto systems, all possible perspectives on beliefs *are already considered as computed.* That is to say, if, in a Toronto system, you want to know what the system believes the user's belief about Frank's belief about the system is, you can simply examine an inner partition of a set of beliefs that has already been constructed and where it is already explicitly stored. This is the exact opposite point of view to that adopted in this paper, which is that such inner environments are not already stored and previously computed, but are constructed when needed and then, as it were, taken apart again, subject to what we call percolation. You can see the difference by asking yourself if you already know what Reagan thinks Begin thinks of Gaddafi. If you think you *already* know without calculation, then you will be inclined to the Toronto view that such inner belief partitions are already constructed. If you think that in some sense, consciously or unconsciously, you have to think it out, you will lean towards a constructivist hypothesis, such as the one advanced in this paper. In the Toronto view, there is no place for least-effort hypothesis of understanding. A fundamental principle of our system is that the points of view are kept, wherever possible, at a single (bottom) level, unless input comes in explicitly stating what person A believes about person or entity B.

The system presented in this paper is in the course of being programmed but, in the form offered here, is essentially a model containing ideas about how to explicate a difficult problem in language understanding. Two things have not concerned us: first, the precise relation of this work to the analysis of speech acts, direct and indirect. These phenomena have been addressed within an AI context (Cohen, Allen, Perrault, q.v.), but we believe a rather different treatment of them will follow naturally—in a later paper—from the prolegomena on belief set out here; one in which the detection of a speech act as being a speech act of particular type will involve more curtailed stereotypical reasoning (speech act reasoning as "frame-like", as it were, rather than plan-or deduction-like), rather than being in the more extended form of those working in the tradition of Searle (1969).

However, we would claim that the belief analysis presented here can be decoupled from this general problem by assuming a parsing into some semantic representation (and the PTs have been a rhetorical rather than a truly technical device in this paper), and passing over the complex problems

of the taxonomy of the inference rules we have assumed. Even if a reader is unwilling to grant this "decoupling" of activities, the proposed computation of nested beliefs can be judged independently, in terms of its psychological and other experimental consequences.

REFERENCES

Allen, J. F., & Perrault, C. R. Participating in dialogues understanding via plan deduction. (AI-MEMO 78-4). Department of Computer Science, University of Toronto, July 1978. (Also *Proceedings of the Second National Conference of the Canadian Society for Computational Studies of Intelligence,* Toronto 1978.)

Bien, J. S. Multiple environments approach to natural language. *American Journal of Computational Linguistics,* 1976 microfiche 54. (a)

Bien, J. S. Computational explanation of intensionality. Reprint No. 41 of the International Conference on Computational Linguistics, Ottawa, 1976. (b)

Bien, J. S. A preliminary study of the linguistic implications of resource control in natural language understanding. ISSCO Working Papers, No. 44, Geneva, 1980.

Bobrow, D. G., & Norman, D. A. Some principles of memory schemata. In D. G. Bobrow & A. M. Collins (Eds.), *Representation and understanding: Studies in cognitive science.* New York: Academic Press, 1975.

Charniak, E. A brief on case. ISSCO Working Papers, No. 22, Geneva, 1975.

Charniak, E. On the use of framed knowledge in language comprehension. *Artificial Intelligence,* 1978 *11,* 225-265.

Clark, H., & Marshall, C. Definite reference and mutual knowledge. In A. Joshi, I. Sag, & B. Webber, (Eds.), *Elements of discourse understanding.* Cambridge: Cambridge University Press, 1981.

Cohen, P. R. On knowing what to say: Planning speech acts. (Technical report No. 118). Department of Computer Science, University of Toronto, 1978.

Donnellan, K. Reference and definite descriptions. In D. Steinberg & L. Jakobovits (Eds.), *Semantics.* Cambridge: Cambridge University Press, 1971.

Gruder, C. Empirical tests of the absolute sleeper effect predicted from the discounting cue hypothesis. *Journal of Personality and Social Psychology,* 1978 *36,* 1061-1074.

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

Moore, R., & Hendrix, G. Computational models of belief and the semantics of belief sentences. SRI Technical Note No. 187, 1979.

Norman, D., & Bobrow, D. Descriptions: On intermediate stage in memory retrieval. *Cognitive Psychology,* 1979, *11,* 107-123.

Putnam, H. The meaning of "meaning". In *Mind, Language and Reality.* Cambridge: Cambridge University Press, 1975.

Searle, J. *Speech Acts.* Cambridge: Cambridge University Press, 1969.

Taylor, G., & Whitehill, S. A belief representation for understanding deception. *Proceedings of the Seventh IJCAI,* Vancouver, 1981.

Wilks, Y. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence,* 1975 *6,* 53-74.

Wilks, Y. Making preferences more active. *Artificial Intelligence,* 1979, *11,* 197-223.

Wilks, Y., & Bien, J. S. Speech acts and multiple environments. *Proceedings of the Sixth IJCAI,* Tokyo, 1979.