[Proceedings of the *Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Colgate University, Hamilton, New York, August 14-16, 1985]

RELEVANCE, POINTS OF VIEW AND DIALOGUE MODELLING

Yorick Wilks

Computing Research Laboratory,
New Mexico State University
Las Cruces, NM, 88003.

1. INTRODUCTION

This paper attempts to compare two approaches to the modelling of human discourse and, more particularly, dialogue. Both place themselves within a general "information processing paradigm", and both descend from the insights of Grice (1975) that understanding is a matter of inference from what is said and what is assumed. So general is that assumption now, and so widespread are the disciplines that draw upon it — philosophy, psychology, linguistics and artificial intelligence (AI) — that it is hard to capture briefly except in opposition to the transformational-generative paradigm of language, with its notions of the primacy and autonomy of syntax, and the theoretical primacy of explications of competence over those of performance. The Generative Semanticists attempted to merge the two traditions and their failure has made it easier to separate off and clarify the work under discussion here.

The two pieces of work to be compared are the work on Relevance Logic by Sperber and Wilson (1982), and that on Points of View and Environments by Wilks and Bien (1979, 1983). The last is a much fuller account than appears here, and the reader is referred to it for more detail. Similarly the present critique of Sperber and Wilson will be found in a fuller form in (Wilks and Cunningham 1984). Both seek to show how one might repair the major lacunae in the work of Grice: exactly what information is to be assumed in the inference processes associated with dialogue, and how it is accessed and manipulated? Sperber and Wilson (SW for short) have made strong claims concerning those, and we shall argue that their claims are misleading or false. We shall then seek to show that our own approach, for all its shortcomings, addresses the problems more directly. One aspect of the comparison we shall offer will be that SWs system remains, at bottom, a process-free linguistic approach, best seen as maintaining certain Chomskyan principles in a plainly pragmatic area, whereas what we offer is a process oriented account, firmly within the AI-psychology tradition.

It may be necessary to make clear that, in making the last distinction, no suggestion is intended that the existence of actual programs is of any theoretical significance in itself. In fact, the work described in the last part of the paper is being programmed but no conclusions are drawn from that. Indeed, it would be wrong to do so because there is much work, in psychology for example, firmly within the information-processing paradigm which does not have programs; as well as work within AI itself where no programs are written, although important points about representation are nevertheless made. In separating SW off, then, from other work in the way we shall do, it is not the existence or absence of programs that is at issue when one refers to the information-processing paradigm. The phrase speech-acts in this paper's title is there only to signal an approach to the outcome from the discussion here: that if "belief spaces", environments, or whatever one calls them, can be located recursively and algorithmically, then the complex manipulations associated with existing work in computation and speech acts (eg Perrault & Allen 1980) will be radically simplified.

Within the broad paradigm of work on discourse understanding as a function of inference and a base of knowledge, belief or assumption, lies a great deal of research in four neighbouring fields, as we noted. We shall greatly narrow that scope here by restricting our attention to work which has some explication of the key notion of an individual's beliefs, and hence some way of distinguishing formally between what one participant in a dialogue believes and what another does, where those two

may be quite different, even though the two people communicate perfectly well. The need for that can be seen very easily: suppose a doctor is talking to a patient and says:

(1) Where in your stomach is that pain, Mr. Smith?

indicating, as he says it, an area of the lower belly. The doctor accepts, by his statement, a lay representation of where the stomach is, i.e. some good way below its actual location. In order to communicate satisfactorily with Mr. Smith he assumes the false belief he believes Mr. Smith to have. He does not himself believe or assert that belief.

Attempts have been made to describe such phenomena within an informal logic of presupposition, but the above is not an example of presupposition on any strict definition of the term and, in any case, the delimitation of that notion is beside the point if what is essential to the example is capturing some notion of "belief space", or partitioning of the assumptions to a discourse, so one can say WHOSE they are.

The need for such a representational facility has long been recognised: in philosophy (e.g. Donnellan 1966), psychology (e.g. Johnson-Laird 1980), linguistics (Levy 1978, Shadbolt 1983) and in AI, particularly in the pioneering work of Perrault and Allen (1980). There has not been such recognition in all AI work on inference and understanding of the last ten years, and our case will be that, whatever the plausibility and sophistication of its claims, there is not that recognition in SW, and it is a fatal drawback.

We shall refer to that representational requirement as "recursive cognitive solipsism" (to use an old philosophical term that Fodor (1980) has deployed recently for other purposes): the requirement that a model of discourse understanding is solipsistic in the sense of modelling only some particular entity's understanding. Hence, the simultaneous modelling of the beliefs of others must be the principal entity's beliefs about those beliefs. Moreover, it must be possible to model the "belief of" operator recursively, to be applied as deeply as is needed for a particular example. In the second section of this paper we offer some suggestions, hopefully original, as to how this might be done.

2. "MUTUAL KNOWLEDGE"

One type of psychological-linguistic account, also descending from the work of Grice, must be mentioned: it is the "mutual knowledge" approach or Schiffer (1972) and more recently Clark and Carlson (1982). This approach accepts much of what has been set out so far and seeks to explicate a formal notion of "A and B mutually know P" as a solution. Although a great deal is idiosyncratically known or believed by individuals, communication is as successful as it is because we not only believe many things in common beings, but believe ourselves to do so. The work of Clark and Carlson thus forms a middle case between SWs "uniform assumption space" approach, which does not attribute ownership of beliefs to individuals (see below for details) and our own attempt to characterise beliefs more generally than by restricting attention to what can be "mutually known". Clark and Carlson's work has been discussed in (Smith ed. 1983) by SW and ourselves, and is not central to the concerns of this paper, but it may be worth recapitulating briefly why that is so.

Their analysis is essentially of situations of actual or potential co-presence: as when two people observe an object lying between them, or go to a cinema together. It is not only that each believes the other to have had such and such an experience, but believes (probably truly) that the other believes that of the first person, and so on..... indefinitely. Infinite numbers of steps like

A believes B believes A believes etc.

are possible, and it is not important for the sake of illustration which of the predicates "believe" or "know" are used here. Clark and Carlson have been misunderstood (not only through the faults of their commentators) as implying that understanders sometimes go through an infinite number of such steps. Of course they do not, and Clark and Carlson did not intend that, but only to claim that to know "A and B mutually know P" is to have the ability to apply a rule as often as is required in a particular case.

We believe that only very special cases can fall under this description, and that not much communication requires assumptions about real copresence. For all other cases, belief is a matter of cognitive solipsism: I truly believe you believe the world is round (rather than flat), just as I do. But this is MY belief not YOURs, and I am unlikely to have or have ever had any direct evidence of the matter. Even in cases of true copresence, as when two people went together to a cinema the night before and refer to the matter later in conversation, the ability to infer infinitely many propositions (as above) is either vacuous (in that there is a rule producing as many as are required for a case, and no evidence is relevant) OR in real cases, evidence is relevant at each stage, and may fail (e.g. on the sixth recursive application of the rule the evidence fails and it is not the case that A believes B believes .....). It is hard to see real inference is much aided by the application of the trivial rule, though it might be handy to have available for special cases, even though the mutual knowledge literature gives no clear guide as to which situations allow its application.

But in the case of real, fallible, applications, storage and effort considerations make it almost impossible that a truth value can reverse after a small number of consistent results (Steele, 1981). In sum then, we find Clark and Carlson's analysis highly ingenious for extraordinary cases but little help in the everyday solipsistic world, where we have no guarantees of what it is both people know, and know each other to know, outside of intensive psychological experiment. As we shall see, understanding does require assumptions about the knowledge of the other partner but they are always fallible, and hence, ex hypothesi, are not "mutual knowledge".

2.1. SPERBER & WILSON'S RELEVANCE ACCOUNT OF UNDERSTANDING

In recent years Sperber & Wilson (SW for short. 1982) have set out a more general account of reasoning than that of "mutual knowledge". They call it a "theory of relevance", and its starting point is Grice's four maxims of communication (1975): they argue that these four can be reduced to one, that of Relation or "Be relevant", and their theory is intended to give content to that rather bare injunction.

The aim of SWs analysis can be stated as one of making explicit the appropriate inferences so as to show, within a single logical space:

what is said, by a speaker

what additional implicit items of information must be brought to bear by a hearer

what inferences follow from the above, including those Grice would have called implicatures

A key term for SW is "contextual implications": these are non-trivial inferences that can be drawn from context and utterance combined, for: "having contextual implications in a given context is a necessary and sufficient condition for relevance" (ibid. p. 73) and deriving the contextual implications is, in effect, the establishment of the relevance of an utterance. A key requirement will be a procedure for establishing what the context (in the sense of a set of propositions as input to an inference procedure) is for a given utterance.

Things get much more interesting when claims move to a quantitative stage and specify the inferences appropriate for understanding in terms of the processing resources available, or offer quantitative selection of the MOST appropriate inference or inferences. This has been done within AI/Psychology under the term "resource limited processing" (e.g. Norman & Bobrow 1975) and, as a special case within the field of natural language processing as "least effort" or "preference" theories (e.g.Bien 1980, Wilks 1975).

SWs version starts with their Principle of Relevance (ibid. p.75), which is "the single principle governing every aspect of comprehension":

"The speaker tries to express the proposition which is the most relevant one possible to the hearer"

This is a counsel of perfection, of course, and may, like all such principles, not be adhered to by the speaker. Here we shall understand it in reverse, as it were, in keeping with the hearer-oriented aspect of SW , as a principle that the hearer is well advised to believe the speaker is observing. But the reader should note in passing that this is not a trivial gloss since this principle, unlike

SWs treatment of the mad-passer-by example, refers only to the speaker's intentions. It is one of the continuing themes of this paper that such perspective switching by SW leads to muddle throughout.

The content of the principle is on p. 75 and we shall call it the Claim:

"Of two utterances that take the same amount of processing, it is the one with the most contextual implications that will be the more relevant; and, of two utterances which have the same number of contextual implications, it is the one which takes the least amount of processing that will be the more relevant".

Some care is needed now, in interpreting this in terms of processing by a hearer, (and SW seem to intend this: "...the hearer has to supply...."(p.73)), since "relevant" in the Claim and the principle may not refer to the same items. This is a problem for SW, one which we will not attempt to solve comprehensively for them here. The real origin of the problem is the ultimate incompatibility of a Gricean speaker's-intention approach and one based on hearer's-information-processing, let alone with an abstract non-directional model based on notions of Chomskyan competence, although SW preserve elements of all these.

2.2. THE THALASSEMIA EXAMPLE

Let us set out their principal illustrative example for their case: the thalassemia example (p.74-5):

"...compare utterances (19)-(21) in a context consisting of (22a-c):

(19) Susan, who has thalassemia, is getting married to Bill.

(20) Susan is getting married to Bill, who has thalassemia

(21) Susan, who has thalassemia, is getting married to Bill, and 1967 was a very good year for Bordeaux wine.

(22)a. People who are getting married should consult a doctor about the possible hereditary risks to their children.

b. Two people both of whom have thalassemia should be warned against having children.

c. Susan has thalassemia.

In this context both (19) and (20) carry the contextual implication that Susan and Bill should consult a doctor, but (20) also carries the implication that Susan and Bill should be warned against having children. The sentences in (19) and (20) are almost identical in linguistic and lexical structure. Suppose that processing involves identifying the propositions expressed by the utterance, computing its non-trivial implications, and matching each of these against the propositions in the context to see if further non-trivial implications can be derived. Then (19) and (20) should take roughly equal amounts processing. In this context since (20) yields more contextual implications than (19), with the same amount of processing, it should be more relevant than (19) and this seems intuitively correct. By contrast, (19) and (21) have the single contextual implication that Susan and Bill should consult a doctor. (21) is linguistically more complex than (19). On the above assumptions about processing, (21) will thus require more processing and be predicted as as less relevant in context; again, this prediction seems to be intuitively correct".

Let us put four immediate considerations against this:

a) What serious quantitative information processing comparison can be going on in which a hearer can be considered as comparing the relevance of two DIFFERENT UTTERANCES (outside explicit psychological laboratory tests, that is) as distinct from realistic situation where a hearer compares two alternative interpretations of a SINGLE utterance so as to select the more relevant? A hearer is normally offered an utterance, not several between which to choose, so what consequences for the information processor could such a formulation propose?

b) It is true, as they note, that (20) produces the, undoubtedly nontrivial, implication that the couple should consult a doctor, but surely "that must have required a great deal of processing to obtain: the location and application of an AND rule, and the location and application of some form of modus ponens to (20) + (22b) AND (22c)? Or do SW somehow imagine that the actual inference itself, normally set out as explicit steps, does not require processing effort? If they believe that they should not use the metaphor at all, but leave it in the hands of others.

If the processing required by inferencing is taken into account, as they seem to intend, then the assumption of equal processing effort required by (19) and (20) is plainly ludicrous, since the accessing of a rule of conjunction and its application is a clear quantifiable cost. In general, the safe assumption, other matters being equal (which, of course, they are not), is that more implications will require more processing effort, exactly the opposite of what the Claim suggests.

c) The real issue has been ignored till now: what basis can there possibly be for assuming, as SW still do at this point in their account, that (19) (20) and (21) all access the same context, and that that access will require the same effort in all three cases (for, if it does not, then the comparisons drawn so far fall to pieces)?

Since (22c) is already present explicitly as part of (19), the context invoked by (19) cannot include (22c) as it does here, (or, if it does, then other parts of utterances can occur explicitly in contexts, which will have other disastrous consequences for SWs Claim). Hence the assumption that (19) and (20) "require the same processing effort" will be quite false if that effort includes context-location (and in the next section they concede that it does). If the effort is not the same then we have another case where the Claim fails to apply, since neither equality is satisfied, although that, too, can give no comfort to SW.

Again, (21) with its mention of Burgundy must draw into the context propositions about wine? They can only be kept out by the indefensible assumption that this simply IS the context, achieved cost free, and declining to discuss the matter further. If the mention of Burgundy did draw in that wider context, at correspondingly greater effort than those drawn in by (19) and (20), then again the whole comparative farrago would fall to bits, since the assumption that (21) will yield the same number of contextual implications as (19) may well turn out false. The same ingenuity that the flag-seller showed with what he heard could certainly produce a context for, and a reply to, (21).

Gazdar and Good (1980) have pointed out that if a hearer has additional or idiosyncratic information about, or interest in, a topic mentioned, then this may well give rise to a great number of non-trivial implications. Had the speaker said, in place of (21), "Susan, who has thalassemia, is getting married to Bill, who is a wine expert", then a wine expert hearer could have correctly inferred a great deal about Bill. SW, lacking any clear notion of what a hearer or speaker believe, separately or about each other, have no defence to this and make none in their reply to Gazdar and Good in (Smith ed. 1982). In a properly founded theory, of course, it would be a requirement that the inferencing was constrained to a sub-space of assumptions that was the hearer's view of what the speaker believed the hearer believed (and the proposals of the last part of the paper are intended to be such). That would meet Gazdar and Good's point, for a hearer behaving appropriately would then not draw such "expertise" inferences if he believed the speaker did not know he was in possession of such information. Only in that way can context finding deal with apparently irrelevant input (and real errors by speakers and hearers can of course occur concerning such information.)

d) The conclusions drawn by SW depend crucially on the context containing the particular premises cited and not others, equally likely to be postulated by a hearer. (22b') is as plausible to a layperson as (22b), and (22d) may be widely believed by normal selfish individuals (and in their paper SW produce a context with a similar selfish belief about the function of charities):

(22b') Thalassemia is a form of bone cancer

(22d) It is unwise to marry a woman with diagnosed cancer, she would need too much attention.

From the whole context (22a 22b' 22c and 22d) (19) now has the implications that they should consult a doctor and that it is unwise of him to marry her. Sentence 20 also has the same two implications from that context, whereas with SW's context (20) carried one more implication. It is obvious that the number depends crucially on the context located and nothing general or significant follows, as SW believe, from the particular examples they chose.

2.3. LOCATING THE CONTEXT

At this point in the exposition, SW introduce a principle that draws the whole theory closer to reality, but has the effect of vitiating much of what has gone before. They now face up to the consequences of the fact that locating contexts is a matter of processing effort:

"We want to argue .....that the search for the interpretation on which an utterance will be most relevant involves search for the context which will make this interpretation possible. In other words, determination of the context is not a prerequisite to the comprehension process, but a part of it. "(ibid. p. 76)

But are we faced at this point by the withdrawal of a minor simplifying assumption, one which can now be withdrawn without ill-effect, or is it rather that the recognition that context finding costs processing effort (one accepted ab initio by all those in the AI-psychology tradition who have discussed the issue) makes nonsense of the Claim and everything based on it?

The following definitions would at least remove the obvious absurdities in SWs position:

New principles for hearers (and assumed by speakers to be in use by hearers)

(1) MAXIMIZE the number of contextual implications drawn for some total given processing effort (up to some arbitrary maximum per unit time, say, one to be empirically determined) for interpretation of input, location of context, and drawing of implications;

(2) MINIMISE amount of processing for context finding so as to leave more available for drawing contextual implications under (1).

These proposals, too, may well not stand up to any detailed examination, but they are at least procedurally plausible and not self-evidently self-contradictory (though they are not independent). Whereas, on SWs account, the hearer is under an injunction simultaneously to maximise and minimise the same sort of thing, in that the hearer is to minimise effort OVERALL, while at the same time maximising the number of contextual implications, whose production must require effort! Our hunch, for what it is worth, is that SW should go simply for a least-processing-effort theory (as we have ourselves) for it is not clear what help having more contextual implications is. As we saw with the flag seller, the number can vary unpredictably with contexts, however chosen, and we believe their addiction to the notion comes from a false view of maximising information in communication.

In their reply to Gazdar and Good (1982), SW deny that they assume "processing speed is constant" (ibid. p.106) and so, they argue, processing per unit time considerations are not relevant. They also claim that context-finding processing is non-inferential, as distinct from the inferential processing that draws implications, and so no considerations drawn from summing processing capacities are appropriate in discussion of their system. This is, we believe, the merest obfuscation, and nothing SW write gives any support to the view that there are separate processing capacities that cannot be added. It is certainly possible that the brain does have separate capacities for the two processes, ones that cannot be added, but claiming that would require some shred of physiological evidence, above and beyond the undoubted convenience to SW. Moreover, such a concession would be quite at variance with the discovery that "determination of context is not a prerequisite to the comprehension process but a part of it (ibid. p. 76)".

Given (1) and (2) above, some version of SWs Claim could now be reinstated, and they might well feel that these new principles are just what they intended and have expressed (cf. "....we would

suggest that the amount of processing tends to remain roughly constant throughout a stretch of discourse" (p.77), if that is taken to mean processing per unit time!), but they certainly have not, and indeed are unable to do so, because the Claim was stated on the basis of the false assumptions about cost-free but which is a matter of complete irrelevance to logical complexity and inferential effort. It would be hard to find in the recent linguistic literature a clearer example of the bad effects of the hangover of beliefs in the autonomy and primacy of syntax.

2.4. COGNITIVE SOLIPSISM

We drew attention earlier to the fact that SW have no clear or consistent appreciation of the fact that real inference must go on somewhere and, in a model of human communication, that must be in a hearer model or a speaker model (where each may, and must, contain models of the other). This lack surfaces in their paper at intervals, as when they discuss the "common ground", which is their version of "mutual knowledge": the set of facts, serving as potential contexts, that both conversational participants know.

But, on SW's account, there is no method or theory to explain the difference between:

(a) premisses/beliefs believed by the hearer to be held by the speaker when speaking.

(b) general knowledge known by the hearer and believed by him to be imputed to him by the speaker

(c) enthymemic beliefs, not retrieved by the hearer from anywhere, but constructed by him, as part of a context, and imputed by him to the speaker on the basis only of what is said (i.e. not believed previously by the hearer, nor previously believed by him to be held by the speaker).

Enthymemic constructions are beliefs attributed to the speaker so as to make the implications follow from what is already in the inference space, e.g. the current utterance. In writing this paper, we produced, out of the blue, more or less plausible sets (22b) and (22d') but as to how this process can be modelled algorithmically no one has much to say. It is the great problem in SWs theory and for all those working in this area. No theory that fails to reflect these distinctions in some well-motivated way can be taken seriously in this area.

Investigations that do make them are the very stuff of much recent work in AI, both logical-theoretical (e.g. Moore 1975) and programmed-applied (e.g. Perrault and Allen 1980). In the important latter series of papers the notions of belief, speech acts, plans, reference, inference, and models of the other, were ingeniously explored, and the above distinctions were fundamental. A crucial advantage of work like the latter type is that it can show the contiguity of relevance and inference with human plans and goals, with what it is someone is talking FOR. There is no way this can be done with SW's work, and that must be a serious shortcoming in a general theory of pragmatics.

More relevant to the matter in hand is the series of papers by Wilks and Bien (1979, 1983) now being programmed, on the procedural location and manipulation of belief spaces or environments in which such inferences can go on, and which simultaneously constrain relevance, inference and the beliefs of the other.

This work is very elementary as yet, and nothing whatever follows from it, but it makes none of the elementary errors of SW. In such a system, one can actually compute (albeit by naive algorithms) the appropriate environment for inference and we shall now turn to more detailed description of that.

Moreover, this work is set within a general claim about least-effort human processing of language, and goes right back to early work (Wilks 1975) on inference chaining in utterance interpretation, in which programs interpreted utterances by establishing the shortest possible chain of inferences from context to utterance. This, as we said, was elementary stuff, but at least it was not incoherent: it was a wholly plausible assumption that locating and applying a shorter chain of inference would require less processing effort.

The greatest lacuna in the list above is (c) and it is, at the moment, no more than an aspiration for all research workers, but SW do not see this because they have no procedural grasp of what can and cannot be done, so for them the very difficult is all one with the well understoood.

Let us summarise our objection to SW's theory: their errors all stem from their conviction that there is an objectively right context set of propositions, one that can be assumed independently of how it is located, and independently of what individuals may in fact believe. Let us now summarise an alternative account of these matter in which those features are built in from the very beginning.

3. AN ALTERNATIVE ACCOUNT: POINTS OF VIEW AND BELIEF ENVIRONMENTS

3.1. INTRODUCTION.

The following dialogue is perfectly natural, and is assumed to take place between a (human) USER and a SYSTEM, that may or may not be human:

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 (II), but you think he does and that is what is important now.

It is clear that, to understand 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 common enough to deserve special attention.

We want to tackle the issue generally (as a representational problem) 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 here will be that there can be a very general algorithm for the construction of beliefs about beliefs or, if you wish, models of models of models, or points of view of points of view of points of view.

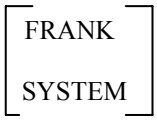
The system has produced replies at points marked (I) and (II) in the dialogue above, and the initial question we ask is why should the system say these different things at these different times, and what structure of knowledge, inference and beliefs about the User and Frank should be postulated in order to produce a dialogue of this type? We shall describe this by saying 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. One could assimilate our account to the most general form of description and say that the system in the dialogue above is interpreting the same utterance (namely, the user's first utterance) differently in the two contexts I and II, where the two contexts are, respectively, the belief space or environment of the system's view of the user's view of Frank's view of himself (the system) at I; and at II the space or environment of the system's view of Frank's view of the system.

The essence of what we shall describe will be how one can express and manipulate such contexts, expressed as in SW as sets of propositions, but now corresponding to the beliefs of individuals and their beliefs about each others' beliefs.

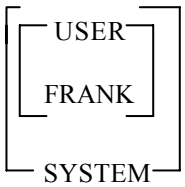
These belief spaces or environments are temporary structures, created in real time during human communication, and not maintained permanently. The reason for that is both a processing and storage one: there would be no argument for storing A's view of B permanently after computing it, unless it was striking and important or something we knew we would need again in the near future. Such structures would be too vast and too many to keep around in any kind of permanent memory as a general feature. We shall suggest that any natural language processing system commited to a least-possible-effort view of the process should act in the way described.

The system described has been developed (Wilks and Bien 1979, 1983) with a view to connecting at some point with the literature on speech act phenomena , and so the computational treatment of the example above would be expected to lead to a discussion of whether anything concerning the notion of a warning was needed to model such a dialogue. Here, however we shall direct the form of our examples towards the type chosen by SW, so that points of contrast can be made.

However, we can use the above dialogue in order to illustrate the shorthand we shall use in the discussion.

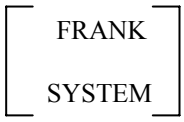


will be used to represent in a schematic form what the bearer of the outer name believes about the bearer of the inner name, that is to say what the system believes about Frank. Structures like this can be nested so that the following structure



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

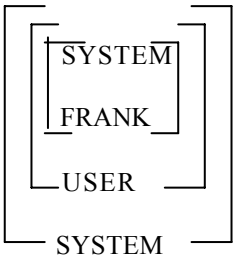
We shall refer to this as a nested environment and every such structure is considered to be (trivially) inside the system, for it knows everything there is to be known about the individuals mentioned. The first important question is, what are the structures that this shorthand represents? For the moment, the simplest form of what the system believes about Frank, i.e.



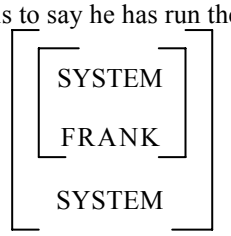
could simply be thought of as less permanent version of a *frame* (Minsky, 1975, Charniak, 1978) or more suitably in the terms of (Wilks, 1977) as a *pseudo-text* or, if you prefer, any knowledge structure whatever about the individual named inside. In this brief discussion we shall refer to them as pseudo-texts (PTs), but since we avoid all questions of semantic representation, they can be thought of as sets of sentences or propositions just like SWs contexts.

The essence of our method is to evaluate and compare two perspectives or environments (i.e. nested PTs) and they will be the ones which are created by the system at points I and II in the dialogue above. "Evaluate" here is intended to have a standard computer science meaning, one we could put more adventurously as *running structural descriptions in given environments* (Wilks and Bien 1979). What this will mean in concrete terms is to draw plausible pragmatic inferences.

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:



Whereas, at Point (II) in the conversation that 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:



where he discovers that he has no such information on what Frank thinks of him. In doing this, at II, the system takes no account of the speaker/users beliefs or motives.

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 these are standard. When we consider nested environments, PT's for agents will be the only ones that can be outer environments in nesting diagrams, because we can consider , for example, Jim's view of the oil crisis but we cannot consider the [...] The principal algorithm computes the set of sentences corresponding to, or included in, a particular nested environment. The form of this can be put very simply: what is in an "inner PT" survives unless explicitly contradicted by an "outer PT". Thus, what I believe A believes about B will be what I believe about B (the "inner PT") except where the "outer PT" in a nesting (in this case, what I believe about A and his beliefs about B) differs. I may believe B is a thief, but I know A does not, hence the final inner resulting set of sentences will not state B to be thief, and will thus differ from my own beliefs.

The process corresponding to the above we describe metaphorically in two ways: a "push down" of one PT inside another, corresponding to the nesting diagrams above, and the "percolation" of beliefs from an outer to an inner PT environment. In practice, that consists in computing the union of the beliefs in the inner PT and the relevant ones in outer PT, and then reducing this to a consistent set by, in the case of contradiction, including only the outer belief in the result. This is a recursive process and can be applied to a nesting of any depth.

The key word "relevant" is interpreted as follows: the beliefs in the outer PT deemed relevant are those explicitly containing the name of the inner PT: i.e. those beliefs in the PT for A containing the name B, in the above example. This is a highly oversimplified method, particularly because it ignores the issues of the identity of individual names and descriptions and whether or not such identities are known by the participants. But, primitive as it is, this "relevance heuristic" is just that, rather than the mystical retrieval procedure envisaged by SW.

The issue of identity of individuals within knowledge and belief contexts is, of course, a central issue in intensional logic, and a method for dealing with it more fully has been developed by (Maida 1983).

Our rule above is a particular case of what is normally called "default reasoning", and the reader is referred to (Wilks and Bien 1983) for a detailed working through of an example of this rule.

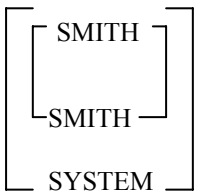
As we noted above, the process assumes not only the default rule but also that belief PTs are *normally* stored only at the bottom level: one does not seek to compute and store full representations of A's beliefs about B, but only beliefs concerning A and concerning B.

Before proceeding to the core of this part of the. paper, a detailed example of the same type as SW's, but treated within the system set out here, we should mention briefly, and in turn, limitations to this "bottom level" assumption and to the "default rule".

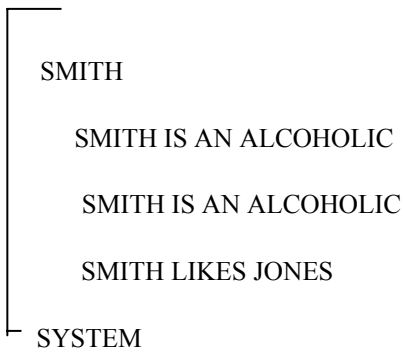
3.2. LIMITATIONS ON BOTTOM LEVEL BELIEFS: BELIEFS OF AND ABOUT.

There is already some partitioning in the PTs, corresponding to the distinction between someone's beliefs *about someone and his beliefs about the beliefs of that individual.* To put it simply, we can have beliefs about Smith, that he is male, 45, etc. etc. We can also have beliefs about his beliefs: that Smith believes that, say, Vitamin C cures colds. On one general view of belief these are all properties of Smith, but they are, of course, importantly different sorts of property.

By representing both these sets of beliefs within the system's PT for Smith, we violate the pure "bottom level" view, for the representation of Smith's own beliefs is already within the nesting



As a form of diagrammatic representation here we shall put a horizontal line across a PT, with the beliefs above the line being those the system believes the individual to hold, while those below are the systems beliefs about Smith, with no commitment as to whether or not he holds them. Thus:



is not a redundant representation.

Moreover, in actual processes we shall make use of "promotion heuristics" for certain predicates like LIKE but not for those like IS-AN-ALCOHOLIC, where SMITH LIKES JONES as a lower belief could be promoted to the corresponding upper half of the PT containing it on the grounds that individuals invariably know whom they like.

One could express this "promotion heuristic" as a set of rules like:

X DISLIKES Y = = = > X BELIEVES (X DISLIKES Y)

A natural additional inference (for inserting incoming dialogue into PTs) would be

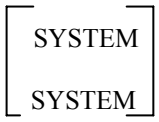
X ASSERT P = = = > X BELIEVE P

unless there was any indication of, or reason to suspect, lying by X. Let us now turn to limitations on the principle of default belief.

3.3. LIMITATIONS ON DEFAULT BELIEF: EXPERT AND SELF KNOWLEDGE.

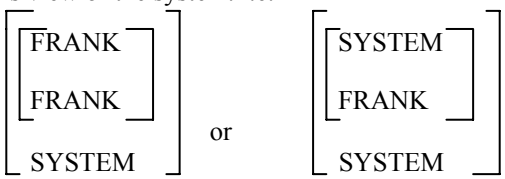
A topic that has not yet been confronted is that of an individual's view of himself, which does not conform to our general heuristics for the computation of points of view: someone else's view of X is my view *except where I believe that not to be the case.*

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



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. In fact, all information in the whole system is (trivially) indexed from this PT.

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

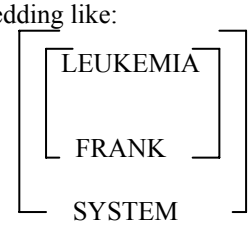


It might be reasonable to assume that Frank's view of himself is, *in general,* the same as my view of him except where I have evidence to the contrary. The default heuristic would create an environment in which Frank believes his address (which I happen to know) to be what I believe it to be; his number of eyes to be what I believe it to be; but his number of teeth must be the value of what he believes it to be (and not the unevaluable function that I have) and so on.

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 exceptions that I am aware of). The normal method of treatment here would be a lambda expression that the system cannot in general evaluate, e.g.

(the (lambda (x)(cardinality-of-Frank's-teeth x)))

The situation described so far is not different in principle from that of expert knowledge, for each individual is an expert concerning himself. So, the general analysis we have given of Frank's "self-embedding" may well hold for any other case of expert knowledge that the system does not share. Suppose Frank is a doctor, but the system has no medical expertise. It may know leukemia is a disease, not its nature, diagnosis or cure, but yet believe that Frank does know these things. Hence in an embedding like:



an application of the default rule will not yield much content but will not be misleading. The inner environment there can be seen as a general "disease PT" or frame, in which there are many empty slots (or, alternatively, as a large number of lambda expressions, unevaluable by the system, of the form (knows(the(x) cure-of etc.)

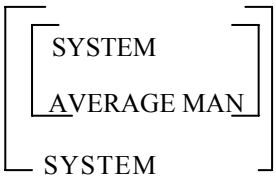
So, in situations where the system represents an average man, or non-expert, constructing the belief environment of an expert by the default rule will not mislead if the system can contains the appropriate PT in a general form. Naturally, it remains a possibility that the environment so constructed is, in reality, hopelessly wrong as to the facts: the system might believe leukemia is a psychiatric condition, for example.

Let us turn to the second case: the computation of Frank's view of the system. On one view, the general default heuristic must surely break down here, because the system cannot assume that Frank has access to all the beliefs about itself that it has. If we applied the general heuristic to a

system that believed itself to be human and with twenty teeth, it would construct an inner environment in which Frank also would have the system's own beliefs about its number of teeth, which is not plausible. One simply knows oneself better than others do.

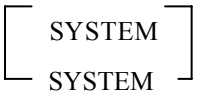
On the other hand, the method of unvaluable expressions used in the first case ought to be equally applicable here, and it ought to be possible for the system to ask and answer the question: "What is Frank's view of me?", if only because this is a common but answerable question in everyday life.

One natural 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 average man's view of it) That would be the entity:



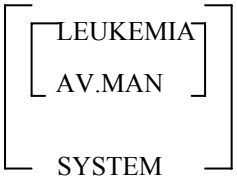
but that, while perhaps solving the defect in the default heuristic, does so at the expense of yet more belief partitioning: more beliefs in subsets whose immediate believer was not the system itself.

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



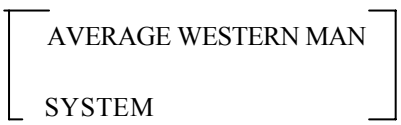
and for them to be the system's beliefs about the average man's view of the system's self. If these were present then the general heuristic would run properly. Clearly the bottom-half beliefs will not necessarily be any subset of those in the corresponding upper half (the system's actual beliefs about itself).

As before, we can look at this issue more generally as one of expertise and consider the case inverse to the earlier one, namely where the system has expert knowledge that the average man does not, as in (1) if the system had medical expertise. The system should then have a belief set corresponding to the environment:



and the set of beliefs so represented would not necessarily be any computable subset of the system's own (expert) beliefs about the disease. In this case no solution is available in terms of a "lower half" of beliefs about an inanimate entity (since the upper-lower distinction cannot be made for non-believers). Hence a general solution must be the first of the two considered above for the system's view of its own public self: namely, the storing of the average man's views, not only of the system itself, but of all areas for which the system considers itself to have expert knowledge, and for which there is a plausible public "non-expert" view.

This could of course lead to a great proliferation of PTs that were not at the "bottom level" in our earlier terminology (i.e. did not have the system as the bottom level believer), and hence to an increase in permanent belief set partitioning. Some economy would be gained from the ability store these hierarchically as for PTs in general (Wilks, 1977), under such nodes as



and, where topics such as semantic definition were concerned the possibility of doing that would correspond precisely to what Putnam (1975) called the "division of linguistic labour" between experts and average men.

3.4. ANOTHER POINT OF VIEW OF MEDICAL EXAMPLES.

Let us now set out as briefly as possible our principal example, making use of the above techniques, and retaining the spirit, though not at all the letter , of SW's thalassemia example. The personae will be a doctor who is hopefully an expert, and of whom we will suppose the system to be a model. There will be a fiancee, who knows the broad facts about the disease, and a fiance who does not, beyond that it is a disease.

The doctor believes them both to carry the disease, and is aware of their different degrees of informedness. Let us also suppose they meet the doctor separately on this matter, and that the following dialogue takes place with the male partner:

(2) He: My fiancee believes I have thalassemia

Doctor: You do and so does she.

at which point the doctor knows that nothing important (i.e. about children) follows for the patient and he must now break the bad news in detail. Whereas, had he been talking to the other partner:

(3) Doctor: You have thalassemia

She: Does he have it too?

Doctor: Yes.

At which point, the doctor knows exactly what follows for the patient and can behave appropriately. It cannot be argued that these distinctions are unimportant, at least it cannot by anyone who subscribes to any degree to the fundamental Gricean insights about human communication. But, as we have seen in the first part of the paper, SW cannot possibly distinguish these situations with their limited theoretical mechanisms.

It will be pretty clear how the environments we have described (contexts in SWs sense) can be constructed to correspond to the doctor/system's views of his two patients. Using the notations above we can suppose the following PT structures:

(4) THALASSEMIA

Thalassemia is a genetic disorder

two carriers should be warned against having children etc.

SYSTEM

(5) THALASSEMIA

Thalassemia is a disease

AVERAGE MAN (we shall write this PT as THALASSEMIA*)

(6) SHE .

She has thalassemia

He has Thalassemia

Intends to marry him and have children

Has expert knowledge about thalassemia

She has thalassemia

She believes he knows nothing about the disease;

SYSTEM

(7) HE

She believes he has thalassemia

Intends to marry her and have children

She knows about thalassemia

He has thalassemia

He is average man about disease

He has thalassemia

SYSTEM

We can further suppose that "he" and "she" as variables cause no problems in this limited world and that the states of the PTs above are after the dialogues (2-5). We can then consider what follows in the following environments (that the doctor/system constructs by push-down of PTs at the appropriate moment), using ordinary inference rules:

(8) THALASSEMIA

SHE

SYSTEM

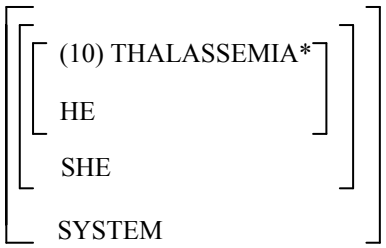
they should not have children, they both have the same disease.

(9) THALASSEMIA

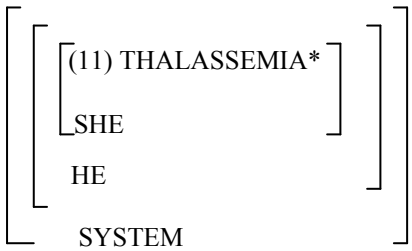
HE

SYSTEM

they both have they both have same disease, she has expert knowledge of the disease they both have.



nothing



This environment contains "He has thalassemia" but no additional consequences follow. Notice that this nesting uses THALASSEMIA*, the average man's view of the disease, because although she does know about the disease, he does not know what it is she does know (though we should really add some unevaluable cure- and diagnosis- functions here from his general PT for disease). Even after:

(12) He: Does she know she has it too?

Doctor: Yes.

nothing beyond the conclusion that they both have it follows because he cannot compute the key fact about the disease that she knows.

It is the main submission of this paper that the above method is rather more realistic that SWs proposals, and it is worth noting that it has nothing at all to do with the number of propositions that follow within any context/environment. Only some method like this one can explain how the doctor treats the patients differently in a way that no impersonal "heuristic of relevance" could possibly do alone.

4. THE CORRECT ROLE OF 'MUTUAL KNOWLEDGE"

An issue discussed at the beginning of the paper, that of "mutual knowledge", can now be seen to be taken care of automatically by the procedures suggested here, and without the need for any special attention. If, in the PT for X's beliefs there is a proposition p, where

p: X and Y both see a candle between them at tn

then when any "push-down" of the PT for Y into that of X is made (to construct the system's beliefs about Xs beliefs about Y and his beliefs) to any required depth, that belief p will always move to the next inner environment, given the crude relevance heuristic proposed earlier.

This transfer of beliefs, without any direct supporting evidence, which we have called percolation of belief, is equivalent to the establishment of a series of propositions:

(the system believes) X believes Y believes p

The iterative percolation, as a result of recursive nesting, will produce, in principle, the infinite belief set found in the "mutual knowledge" literature. In other, related, papers (e.g.Wilks and Bien, 1983) we have emphasised that it is just such recursive (let alone infinitely recursive) pushdowns that any system based on a Principle of Least Effort will seek to minimise.

Whereas, for a "lower half or "aboutness" belief (believed by the system), such as

X and Y are mafiosi

we would expect promotion to the inner belief set, and so as a candidate for iterative percolation, only if the predicate were appropriate (and IS-A-MAFIOSO presumably is such a predicate since if you are one you know it!), but then only for the half of the conjunction drawn in by relevance: e.g. there is no reason why Ys being a mafioso should be promoted to the self-beliefs of X, just because that is entertained as a conjunction by the system. It would clearly be a delicate issue to settle just which predicates were so "promotable" but that is equally a problem for Clark and his collaborators.

But given any such typing of predicates, the general inference rule of the system does the rest without there being any special consideration at all of mutual knowledge phenomena. They have no privileged place, but are just epiphenomena that have arisen in the literature because of an inadequate general characterisation of beliefs about beliefs and their computation.

In (Wilks & Bien,1983) it was argued that some such percolations of unsupported belief into inner environments might be expected to have psychological consequences analogous to the so-called "sleeper effect". It would be interesting if that occurred as the side-effect of the operation of a very general rule for constructing nested environments.

5. CONCLUSION

It is hoped that the reader will be able, from this brief sketch, to assess and compare the two approaches discussed in the paper, and decide which offers a more realistic account of the pragmatic inferences individuals make on the basis of their actual beliefs, including their beliefs about others beliefs.

Finally, some brief note should be made of the differences between this work and similar work on speech acts and plans done at Toronto by Perrault, Allen and Cohen (see Perrault, and Allen, 1978; Cohen, 1978 etc.). One is their emphasis on plans, which are not of central concern here. 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, then you can simply examine an inner partition of a set of beliefs that has already been constructed, 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 stored already, and previously computed, but are constructed when needed and then, as it were, taken apart again, subject to percolated beliefs remaining behind with the effect the nested environment so constructed is lost when the need for it is gone. It is this that keeps belief structures stored as much as possible at what we have been calling "bottom level".

You can see the difference by asking yourself if you already know what Reagan thinks Gorbachev 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, like the one advanced in this paper.

6. REFERENCES

Bien, J.(1980), in PROC. IJCAI83, 675-677.

Charniak, E.(1978) in ARTIFICIAL INTELLIGENCE, Vol.11, pp.225-265.

Clark, H. & Carlson, T., (1982), in Smith (ed.) pp.1-37

Cohen, P. R.(1978) in JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY, Vol.36, 1061-1074.

Donnellan, K. (1966) PHILOSOPHICAL REVIEW. 75, 281-304.

Fodor, J. A. (1980) in BEHAVIORAL AND BRAIN SCIENCES, Vol. 3, 63-73.

Gazdar, G. & Good, D.(1982), in Smith (ed.) pp.88-100

Grice, H. (1975), in Cole & Morgan (eds.) SYNTAX AND SEMANTICS: SPEECH ACTS, Vol.3, London: Academic, pp.41-58.

Johnson-Laird, P. (1980) COGNITIVE SCIENCE, 4, 71-115.

Levinson, S. (1983) PRAGMATICS, Cambridge: CUP.

Levy, D. (1979), in (T.Givon, ed.) SYNTAX AND SEMANTICS, Vol. 12, New York: Academic Press, pp.183-210.

Maida, A. (1983), in PROC. IJCAI83, 179-183.

Minsky, M. (1975) in (P.Winston, ed.) THE PSYCHOLOGY OF COMPUTER VISION, New York, pp.211-277.

Moore, R.(1975) Reasoning from incomplete knowledge in a procedural deduction system, MIT-AI Lab., AI-TR-347.

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

Perrault, R. & Allen, J. (1980) in AMER.JNL.OF COMPUT.LINGUISTICS, Vol.6, 167-182.

Putnam, H.(1975) in MIND LANGUAGE AND REALITY, Cambridge, pp.215-232.

Schiffer, S.(1972) MEANING, Oxford: Clarendon Press.

Shadbolt, N.(1983) in JOURNAL OF SEMANTICS, Vol.2, 63-98.

Smith, N. (ed.)(1982) MUTUAL KNOWLEDGE, London:Academic.

Sperber, D. & Wilson,D. (1982) in Smith (ed.), pp.61-87.

Steele, S.(1981) personal communication.

Wilks, Y. (1975) in ARTIFICIAL INTELLIGENCE, Vol 6, 88-111.

Wilks, Y. (1977) in ARTIFICIAL INTELLIGENCE, Vol.8., 75-97.

Wilks, Y. & Bien, J. (1979) in PROC.IJCAI79, 451-455.

Wilks, Y. & Bien, J. (1983) in COGNITIVE SCIENCE, Vol. 8, 120-146.

Wilks, Y. and Cunningham, C. (1984) "A purported account of semantic relevance" Essex Cognitive Studies Memorandum, no. 16.