Phil. Trans. R. Soc. Lond. B **295**, 353–374 (1981)
Printed in Great Britain

Comprehension as the construction of mental models

By P. N. Johnson-Laird

Centre for Research on Perception and Cognition, Laboratory of Experimental Psychology, University of Sussex, Brighton BN1 9QG, U.K.

This paper presents a theory of how language is understood, and gives some supporting experimental evidence. Its fundamental hypothesis is that discourse is sometimes mentally represented in a form akin to that of perceived or imagined events. Skilled narrators have the power to elicit such representations so that their audiences seem to experience the events rather than merely to read or hear about them. The theory assumes that there are two main stages in comprehension. First, utterances are translated into a mental code that provides a direct linguistic representation of them. This stage concerns the identification of speech sounds, the recognition of words, and the recovery of superficial syntactic structure. Secondly, the linguistic code may be used as part of the basis for the inferential construction of a *mental model* of the state of affairs that the utterances describe. On some occasions, listeners go no further than the first stage of interpretation.

Several lines of research support the two-stage theory. If people construct a mental model of a discourse, then, for example, their memory for its gist is better than if they have failed to do so, but their recall of verbatim detail is poor. If people do not construct a mental model of a discourse but rely solely on the linguistic code, then they tend to remember the overall import of the passage poorly, but they are often able to recall verbatim detail. Such contrasting results were obtained by comparing determinate descriptions with indeterminate ones that could not be accurately represented by a single mental model. The paper presents a number of other phenomena concerning the coherence of discourse that corroborate the theory.

Introduction

Meaning is elusive. You know at once whether or not you understand an utterance, and often whether or not you agree with what it asserts or can comply with what it requests. But, no matter how hard you try, you cannot grasp the form in which meaning is represented or the nature of the mental processes underlying comprehension. This elusiveness made possible the doctrine that there are no such things as meanings or minds – the doctrine of Behaviourism that was defended for several decades by psychologists, linguists and philosophers. This paper, however, will offer a conjecture about both meanings and minds, and advance some evidence to support it. The conjecture is simple. A major function of language is to enable one person to have another's experience of the world by proxy: instead of a direct apprehension of a state of affairs, the listener constructs a model of them based on a speaker's remarks. What needs to be elucidated is the nature of *mental models*, and how they are constructed during comprehension. To set the scene I shall begin with a brief account of how language can be related to models in an ideal universe – the relation that is provided by model-theoretic semantics.

The basic idea of model-theoretic semantics is this: on the one hand, there is a language that is recursively defined by syntactic rules governing the set of well-formed expressions; on the

other hand, there is a model-structure consisting of a domain of individuals, numbers or entities of some sort. A semantics for the language is provided by specifying a function that yields an interpretation with respect to the model structure for each expression in the language. The basic terms of the language are mapped directly onto the appropriate sets of individuals in the model structure, e.g. predicates map onto the sets of individuals possessing the relevant properties. Other rules build up the interpretation of a sentence recursively as a function of the interpretations of its parts. There is accordingly a commonly accepted methodological assumption that these rules work in parallel to the syntactic rules governing well-formedness. They also introduce syncategoramatic words such as *and* and *or*; and their final product is to deliver the truth value of a sentence with respect to the model structure. The use of such methods for the analysis of *natural* language relies on introducing a set or sets of model structures; each model structure represents a possible state of affairs – a 'possible world' – at a particular moment (see Montague 1974). If a theorist wishes to capture the synonymity of two expressions such as *seek* and *try to find*, then it can be stipulated in a so-called 'meaning postulate', which has the effect of eliminating certain otherwise possible worlds – those in which someone could search for something without trying to find it. There are several further layers of complexity that mask the fundamental simplicity of a model-based theory of meaning.

The introduction of possible worlds into the models of formal semantics illuminated Frege's (1892) distinction between sense (intension) and reference (extension). The intension of a sentence – the proposition that it expresses – can be treated as a function from the set of possible worlds onto the set of truth values (true and false); the extension of the sentence is its truth value with respect to the particular 'possible world' under consideration. Hence, the difference between the obvious truism, 'The Morning Star is identical to the Morning Star', and the true but informative assertion, 'The Morning Star is identical to the Evening Star', is that the first is true in all possible worlds, whereas the second is true only in some possible worlds (including the real one). The two assertions accordingly have different intensions but the same extension in our world. As a matter of fact, this account is an oversimplification, because the proposition expressed by a sentence in ordinary language depends on the context in which it occurs: the time and place, the speaker and listener, and the circumstances of the utterance. It is therefore more appropriate to think of one set of possible worlds as representing the context of a sentence's utterance, and to treat the meaning of the sentence as a function from contexts to a proposition, and the proposition in turn as a function from possible worlds to a truth value (see, for example, Stalnaker 1978).

Although model-theoretic semantics provides extremely powerful techniques for the analysis of meanings, it stands altogether outside psychology. It relates a language directly to a model structure, whereas natural language relates to the world by way of the human mind. Model-theoretic analyses are built on the assumption of simple functions that map basic words onto the model structure. The relation between the lexicon of natural language and the world is very much more problematical. Moreover, not everything is for the best in the semantics of 'possible worlds': the treatment of sentences about beliefs, desires and other propositional attitudes or psychological states has yet to be satisfactorily formulated within its framework. You can see intuitively that if language is mapped directly onto a model structure without interposing a mind between them, then there are likely to be difficulties in analysing assertions about the contents of minds. These remarks should not be construed as criticisms of formal semantics: it was not designed as a psychological theory, and it certainly has sufficient power to express any theory of meaning that can be clearly formulated.

Language and mental models

Although model-theoretic semantics stands altogether outside psychology, there are some analogies between it and psychological semantics. One of its standard methodological conveniences is to provide an explicit way of translating expressions in natural language into a formal language, such as a tensed intensional logic, which in turn is interpreted with respect to a model structure. Strictly speaking, there is no need for the intermediary language: it merely facilitates the process of analysis. However, it offers an obvious analogy for the treatment of meaning in psychology. Perhaps utterances in natural language are translated into expressions in a mental language, 'mentalese', which in turn is interpreted with respect to the world. This notion of a mental language for representing meaning has been canvassed in one form or another by everyone who has proposed a psychological theory of meaning (compare, for example, Collins & Quillian 1972; Kintsch 1974; Smith *et al.* 1974; Fodor *et al.* 1975). It naturally suggests that the entailments of a sentence are recovered by using rules of inference that enable new expressions in the language to be derived from old; these rules include both standard logical schemata and 'meaning postulates' that establish the appropriate semantic relations betwen predicates, e.g. for any x, if x is a bachelor than x is a man and x is not married. This idea has been formulated in many different ways, from dictionary entries that decompose meanings in accordance with meaning postulates to semantic networks in computer programs that link words according to meaning postulates.

There are a number of grounds for supposing that at the very least such systems must be supplemented by a richer form of representation, namely mental models of the states of affairs the utterances describe. These grounds have been spelt out elsewhere (see Johnson-Laird 1978, 1980), but, in essence, the need for mental models derives from the necessity of having some machinery for handling the reference of expressions, and for making deduction possible without having to rely on rules of inference. Indeed, in so far as natural language relates to the world it does so through the mind's innate ability to construct models of reality.

The capacity to draw inferences similarly rests on the ability to construct and to manipulate mental models. The errors that occur in reasoning with quantified assertions can be readily explained on the assumption that people do not always search assiduously for alternative models of the premises that render a putative conclusion false. In so far as human beings have internal rules of inference that operate on mentalese representations, they probably derive them from invariant outcomes in the manipulation of models. The origins of formal logic as an intellectual discipline are likely to arise from an awareness of potential error as a result of failing to carry out the search procedures exhaustively, and in a self-conscious attempt to externalize the process.

The notion of a mental model is highly relevant to the comprehension of discourse, especially its referential aspects. A number of theorists have indeed put forward the idea that such interpretations rely on the construction of a discourse model (Karttunen 1976; Stenning 1977; Webber 1978; Johnson-Laird & Garnham 1980). From a psychological standpoint, a discourse model represents the state of affairs described by the discourse, and it is set up as a function of its constituent sentences, and inferences based on a knowledge of context in its widest sense and on general knowledge. The role of inference is crucial, because nearly all utterances depend on context for their interpretation. To put the point another way: it is almost impossible to make a message so explicit that it is not dependent on its context. If you are in doubt about this claim, consider just two aspects of the problem. First, to render temporal reference explicit,

it is necessary to refer to calendars and clocks but, to make sense of such expressions as 10.30 a.m. Greenwich Mean Time, 23 December 1143, it is necessary to know about the way in which time is conventionally recorded. Secondly, any attempt to make the illocutionary force of an utterance wholly explicit can never in itself guarantee that force. Suppose that someone remarks,

I hereby assert to you that Floyd broke the glass

then it is always possible that, in fact, they are making no such assertion: they could be quoting a linguistic example, acting in a play, or playing the fool. The retrieval of the true illocutionary force of an utterance always depends on a knowledge of the context and circumstances that gave rise to it.

Utterances are clues – or, more properly, cues – to the interpretive process. It goes to work on them by engaging both linguistic and inferential processes. A major problem for these constructive processes is the indeterminacy of linguistic descriptions. Just as utterances depend on context for their interpretation, so, too, they almost invariably fail to eliminate a number of alternative interpretations. Let us turn to a more detailed analysis of the issues.

Referential indeterminacy

Here is a description of a person's room:

> I have a very small bedroom with a window overlooking Hampstead Heath. There is a single bed against the wall and opposite to it a small gas fire with a gas ring for boiling a kettle. Between the gas fire and the window at the end of the room, there is a bookcase with all my books in it, and my portable radio standing on top. The door opens directly on to the head of the bed, but on the wall on its right hand side there is a small wardrobe where I keep my clothes. The room is so small and narrow that I cook sitting on the bed and can easily reach over to the gas ring.

Such a description is entirely typical, and, ignoring its lack of literary quality, suffices to create a reasonably clear impression of the room. However, it is radically indeterminate, that is, it is consistent with a potentially infinite number of alternative rooms. The problem is not merely a question of the dimensions of the room and the objects it contains, but rather that the actual layout itself is not described in a way that permits a unique reconstruction of the spatial relations between objects. How, then, is such a passage mentally represented? A number of authors have struggled with this problem, which was first brought home to me on reading Miller's (1979) discussion of it. As he points out, one can analyse the interpretation of discourse within a model-theoretic approach by assuming that each successive sentence reduces the number of alternative 'possible worlds' that are compatible with the passage. In the framework of mental models, however, there are two alternative strategies for coping with radical indeterminacy. First, a listener may simply refrain from continuing to construct a model as soon as its indeterminacy becomes apparent. Indeed, its indeterminacy is only likely to be detected in the process of trying to construct a model. Secondly, a listener may plump for one specific interpretation of the discourse, perhaps on the basis of an educated guess. If the guess is correct, then there is plainly no problem. Only if it proves to be wrong in the light of some subsequent information is there a difficulty. In this case, the listener may either attempt to reconstruct an appropriate alternative interpretation or else be forced to abandon the whole enterprise in some confusion. One observes both of these outcomes when travellers receive detailed instructions

about a route. We have investigated them experimentally and by developing a computer program that implements the reconstructive strategy.

In one of our experiments, K. Mani and I examined the idea that readers can construct a mental model of a spatial layout (Mani & Johnson-Laird 1981). The subjects heard a series of spatial descriptions such as the following example.

The spoon is to the left of the knife
The plate is to the right of the knife
The fork is in the front of the spoon
The cup is in front of the knife

They then judged whether a diagram such as

spoon	knife	plate
fork	cup	

was consistent or inconsistent with the description. If you think of the diagram as depicting the arrangement of the objects on a table top, then obviously it is consistent with the description. Half the descriptions were determinate as in the example, but the other half were grossly indeterminate. The indeterminacy was introduced merely by changing the last word in the second sentence.

The spoon is to the left of the knife
The plate is to the right of the spoon
The fork is in front of the spoon
The cup is in front of the knife

This description is indeterminate in that it is consistent with at least two radically different arrangements.

spoon	knife	plate	spoon	plate	knife
fork	cup		fork		cup

After the subjects had evaluated a whole series of pairs of descriptions and diagrams, which were presented in a random order and each with a different lexical content, they were given an unexpected test of their memory for the descriptions. On each trial, they had to rank four alternative descriptions in terms of their resemblance to the actual description that they had been given. The alternatives consisted of: the original description, a description that was inferable from a model of the original description, and two confusion items that described arrangements different from those in the original description. The inferable description for the example above included the sentence

The fork is to the left of the cup

which could be inferred from the layout corresponding to the description (either the determinate or the indeterminate one).

The subjects remembered the gist of the determinate descriptions very much better than that of the indeterminate descriptions. The percentage of trials on which they ranked the original and the inferable description as closer to the original than they ranked the confusion items was 88% for the determinate descriptions, but it was only 58% for the indeterminate descriptions. All 20 subjects conformed to this trend, and there was no effect of whether or not

a diagram had been consistent with a description. However, the percentage of trials on which the original description was ranked higher than the inferable description was 68% for the determinate descriptions, but it was 88% for the indeterminate descriptions. This difference was also highly reliable.

A plausible explanation for the pattern of results is that the subjects construct a mental model for the determinate descriptions but abandon such a representation in favour of a superficial linguistic one as soon as they encounter an indeterminacy. Mental models are relatively easy to remember but encode little or nothing of the form of the original sentences on which they are based, and subjects accordingly confuse inferable descriptions with the original sentences. Linguistic representations are relatively hard to remember, but they do encode the linguistic form of the sentences in a description.

Attempts to use the reconstructive strategy in which an existing model is modified in the light of subsequent information are likely to place a very much greater load on the cognitive system, and it is therefore hardly surprising, given the nature of the experimental task, that the subjects in our experiments tend to abandon models in favour of linguistic representations. The nature of this additional load can be illustrated by considering a computer program (written in the list-processing language, POP-10). The program builds up a spatial model of the relations between entities, combining the information from separate assertions to produce a composite representation. The program contains a number of general procedures that carry out the following tasks: adding a new item to an existing array on the basis of an assertion that relates it to an item already in the array, initiating the construction of a new array on the basis of an assertion that refers only to items not in any existing array, testing whether a specified relation holds between items on the basis of an assertion that refers only to items in an existing array, and forming a single composite array on the basis of an assertion that relates hitherto unrelated items. The general procedure for verifying whether a specified relation holds between two items works by scanning along a line, whose origin is the second of the two items, to determine whether the first of them is somewhere on that line. The direction of this search is controlled by two variables, which are the values by which the X and Y coordinates are incremented to spell out successively the locations to be scanned. The actual values of X and Y are determined by the spatial relation expressed in the assertion to be verified. Indeed, all the general procedures make use of the same mechanism for specifying the relevant direction. This makes it possible to employ very simple lexical entries for the relational terms recognized by the program.

In fact, the dictionary entries of spatial relations are in a marked contrast to the sorts of representations proposed by orthodox theories of psychological semantics, whether they rely upon decomposition into semantic elements (see, for example, Smith *et al.* 1974), or a semantic network that links words according to the relations between them (see, for example Collins & Quillian 1972), or meaning postulates that specify the semantic relations between words (see, for example, Fodor *et al.* 1975). Instead of a specification of logical properties, the lexical entry for *behind*, for example, simply consists of an instruction to take one of the general procedures, and to 'freeze in' specific values for two of its parameters, in this case the values of 1 and 0. If the general procedure is the one for verifying the relation between items, a new more specific function is created: the parameters 1 and 0 specify the direction in which to scan from the second item to verify whether the first item is behind it, i.e. increment the Y coordinate and hold the X coordinate constant. The viewpoint of the program is accordingly analogous to that of a person examining the array as though it were a graph laid out in front of him on a table.

I have dwelt on these details because they establish a number of relevant theoretical points. First, it is feasible to have a semantic system in which the meanings of words in the object language (in this case a very restricted subset of English) are represented in an internal meta-language that has no simple, let alone one–one, relation to the object language. Such processes as 'freezing in' the values of variables certainly cannot be described in the subset of English that the program comprehends: they are difficult enough to express within the full resources of English. Secondly, the program establishes that it is possible to represent the meanings of relational terms without having to specify their logical properties explicitly. It is not necessary to state, for example, that the relation *behind* is asymmetric, irreflexive and usually transitive (though to varying degrees). The expression is simply assigned a semantics that makes it possible to build up mental models from which these logical characteristics are emergent properties (Johnson-Laird 1975, 1978). For example, if the program is told, 'A is behind B', then it constructs the following array.

A
B

If the program is then told 'B is behind C', it adds the new information to the existing array.

A
B
C

If, at this point the program is told, 'A is behind C', then on finding that both A and C are in the existing array, the program runs its verification procedure: it looks to see whether A is indeed behind C, and, since it is, it returns the value 'true'. In short, it makes a transitive inference without having to possess an explicit representation of the transitivity of *behind*. Transitivity is an emergent property of the system of interpretation based on spatial arrays and a procedural semantics for lexical items. A more complex procedure is, of course, required in real life: it is necessary to allow that the locus of points scanned need not be a straight line, and that the actual shapes and sizes of objects will affect it. It is also necessary to allow a degree of latitude in the scanning procedure. Nevertheless, it seems that the only way that such factors could be taken into account would be by constructing an internal model of the sort described here.

Theorists such as Jerry Fodor (1975), who are strongly committed to the doctrine of innate ideas, may protest that transitivity is embodied in the machinery for interpreting arrays. They may likewise argue that there is no way in which the logical power of a system could be increased by learning. The first point is false; the second point is ambiguous. Any computable function can be characterized in terms of three sorts of basic function and three operations to create new functions out of old functions (see, for example, Boolos & Jeffrey 1974). The three basic sorts of function are the zero function, which returns the value of 0 for any input argument, the set of identity functions, which return the value of one specified input argument from the set of input arguments, and the successor function, which returns the successor of any integer, e.g. given 3 it returns 4. The three operations are composition, which allows the outputs of one set of functions to provide the input values of another function; recursion, which is a rather more complicated matter that I shall sketch presently; and minimization, which sets the value of one function equal to that value of an argument of another function that is the smallest value for which that function yields zero. The point to be emphasized is that any effective procedure

whatsoever, as far as anyone knows, can be analysed in terms of these functions and operations. Hence, one way in which to think of learning is that the system discovers a new way to combine functions that it already possesses, in order to create a function that it has not hitherto possessed. There is accordingly both a sense in which the system does not increase in power – it always possessed the power of a universal Turing machine – and another sense in which the system *does* increase in power – hitherto it was unable to compute, say, a square root, but now it can do so. In the light of this argument, the question of whether the mechanism for manipulating spatial arrays contains a rule for transitivity can be resolved. An array can be treated as a set of memory locations that are defined by ordered pairs of integers. The semantics of *behind*, for example simply requires a function that takes an initial pair of integers, such as (2, 1) and applies the successor function iteratively to the first member of the pair to produce the following sort of sequence: (2, 1), (3, 1), (4, 1), and so on. At each memory location, the procedure examines the contents to check whether they correspond to the first item referred to in the assertion. Hence, although the system abides by the principle of transitivity, in no sense does it contain a representation of it.

After this lengthy excursion into procedural semantics, let us return to the reconstructive strategy as it is employed by the program. If the program is given a radically indeterminate description such as

A is on the right of B
C is on the left of A

it plumps for just one model, as follows.

C B A

If it is subsequently told that C is on the right of B, then the verification procedure initially returns the value 'false'. However, whenever it has this outcome, the program calls another general procedure that checks whether there is any way in which the model could be reorganized so as to render the assertion true. (In the same way, whenever the verification procedure returns the value 'true', there is a general procedure that checks whether there is any way in which the model could be reorganized so as to render the assertion false.) A moment's thought should convince you that a procedure that follows such a reconstructive strategy must be rather powerful computationally. To revert to the example, the procedure must check whether, if it switches round the positions of B and C, the result is still compatible with the premises. In this particular case, it can do so freely, but suppose that a previous assertion had established that D is in front of B, then the model would be as follows.

C B A
D

In this case, before B can be switched with C, it will be necessary to check whether the position of the D can similarly be shifted. In fact, D can be shifted, but again, suppose that in still earlier assertions it had been established that E is in front of C, and that E is on the left of D, as follows.

C B A
E D

In this case, there is no way in which the model can be revised so as to render true the assertion, 'C is on the right of B'. In principle, there is no limit to the number of 'dependents' of B and

C that have to be followed up to check on the reconstruction. This aspect of the problem requires the reconstructive procedure to have the computational property of *recursion*, i.e. the procedure must be able to call itself.

The notion of recursion is apt to be puzzling when it is first encountered, and the reader may find it helpful to consider the following example. The concept of being related to someone by ties of kinship can be given a recursive definition:

x is related to y = x is married to y or x is a parent of y;
x is related to y = x is related to z and z is related to y;

where x is a child of y is equivalent to y is a parent of x. Thus, for example, the following chain of inferences may be necessary to establish that Arthur is related to Zoë: Arthur is related to Bill and Bill is related to Zoë (because Bill is married to Cynthia and Cynthia is related to Zoë (because Cynthia is a parent of David and David is related to Zoë (because...[and so on to any arbitrary depth])))). Ultimately, the recursion has to be completely spelt out by appeals to the first line of the definition or else the relationship between Arthur and Zoë will not be defined. Some recursively defined functions can be computed in a simpler way, but other functions can only be computed recursively. The only way to *parse* unlimited self-embedded structures, for example, is by a recursive procedure, and this fact establishes the connection between a particular sort of recursive rule in linguistics and a recursive procedure. However, as we have seen, there are also general cognitive procedures such as the rearrangement of mental models that may also require recursion: it is not peculiar to linguistic processing.

Comprehension and veridicality

Because language is used to create worlds by proxy, it does not matter as far as psychological processes are concerned, whether these worlds are veridical. The processes by which fictitious discourse is produced or understood are not strikingly different from those for true assertions. That, perhaps, is why language is so powerful a medium for creating beliefs. You construct a mental model in much the same way whether you are reading *War and peace*, or an obvious work of fiction such as your daily newspaper. It follows that reference and the mechanisms that underlie it do not depend on whether there is something in the world that an expression picks out. What *is* crucial is that there should be something in a mental model to which the expression refers. This is not to say that questions of veridicality do not matter – obviously it is important to know whether or not there is a present king of France, and, perhaps to a lesser extent, whether God exists. But these are not matters that affect linguistic processing. They do not even seem to affect linguistic usage in many languages, though perhaps there are languages where existence–nonexistence has a morphological reflex.

A simple illustration of the independence of referential usage from questions of existence is provided by one aspect of discourse concerning propositional attitudes. Suppose, for example, I remark:

I'd like to meet the man who lives at 221*b* Baker Street

and my use of the definite description is *attributive* in the sense that Donnellan (1966) has popularized, i.e. I intend to pick out whoever satisfies the description. You, who know perfectly well that there is no one who lives at that address, can report my views:

Phil would like to meet the man who lives at 221*b* Baker Street, whoever he may be

without being committed to the existence of such an individual, except in your model of my beliefs. What is less obvious, however, is that even when referential designations are used in the context of a propositional attitude, there may be a failure of reference that does not affect the truth of the assertion. If, for instance, I share the allegedly common delusion of believing that Sherlock Holmes is a real person, who lives at 221*b* Baker Street and who is the greatest detective in the world, then you, who know all these facts about my beliefs, may instead report my remark in the following words:

Phil would like to meet Sherlock Holmes.

Your designation is *referential* by Donnellan's criteria: in particular, you are committed to the substitution *salva veritate* of other designations that also pick out the same individual in your model of my beliefs, e.g.:

Phil would like to meet the great detective who lives at 221*b* Baker Street.

Yet you do not intend to designate an actual person. Hence, the question of existence is independent of the referential–attributive distinction.

Existence – at least as far as the speaker is concerned – is the critical factor for such inferences as:

The king believes that someone has betrayed him.
Therefore, there is someone whom the king believes has betrayed him.

or:

The king believes that Arthur Ortcutt has betrayed him.
Therefore, there is someone whom the king believes has betrayed him.

An inference that leads to a conclusion in which the quantified noun phrase is outside the scope of the propositional verb is warranted provided the originally designated individual exists. Contrary to what has been assumed by some theorists (e.g. Kaplan 1969), the designation may well be attributive rather than referential. The king and I may believe that he has been betrayed by the man in the iron mask, but we may know nothing else about this individual. I might accordingly assert:

The king believes that the man in the iron mask, whoever he may be, has betrayed him

and thereby intend to be committed to the inference:

There is someone whom the king believes has betrayed him.

If the king merely believes that someone or other is a traitor, then the validity of 'exporting' the quantifier from the context of belief again depends solely on the question of existence.

Granted that the referential–attributive distinction is independent of propositional attitudes, then it follows that any attempt to explain that distinction by reference to particular propositional attitudes is unworkable (*pace* Hintikka (1969)), and likewise that the converse attempt to elucidate referential opacity in terms of the referential–attributive distinction is equally erroneous (*pace* Cole (1978)).

Reference and mental models

There are a number of other referential phenomena that are best made sense of by postulating the existence of mental models of discourse. As Johnson-Laird & Garnham (1980) have argued, if a speakers asserts:

The man who lives next door to me bought a birdbath

then no one would ordinarily take this remark either to entail (*pace* Russell (1905)) or to presuppose (*pace* Strawson (1950)), that there is one and only one man living next door to the speaker. The speaker is simply referring to the only neighbour who is – or who is going to be – relevant in the current context. Johnson-Laird & Garnham (1980, p. 377) write:

> Uniqueness in a model rather than in reality is what controls the use and interpretation of definite descriptions. If a speaker is to communicate felicitously, then he must consider whether an entity will be unique in his listener's model. Utterances need seldom be more than clues about how to change a discourse model: they depend for their interpretation on what a listener knows, but that interpretation in turn modifies or extends the discourse model. A discourse model is in part a surrogate for reality. Indeed, it is sometimes convenient to speak as if language were used to talk about discourse models rather than the world.

As an illustration of this thesis, Garnham and I re-examined the distinction between a referential and an attributive use of a definite description. Donnellan (1966) pointed out that a referential use of a description such as *the murderer of Smith* occurs when a speaker intends the listener to pick out the particular individual about whom he is speaking. Alternatively, an attributive use occurs when the speaker intends the listener merely to pick out whoever satisfies the description. It is natural to suppose, as Donnellan did, that in an attributive use there must be someone who satisfies the description, or else the assertion will be vacuous; but where a description is used referentially, such a failure is not so critical, since a listener may nevertheless be able to discern the speaker's intentions and recover the appropriate referent even though it is misdescribed. Despite the plausibility of this argument, Johnson-Laird & Garnham show that there is a flaw in it.

Suppose that a friend of yours has been to the cinema, then you might ask his wife,

How did John enjoy the film he saw last night?

where your usage is plainly attributive since you know nothing about the film and can designate it in no other way. Similarly, John's wife might reply,

He enjoyed it

using the pronoun attributively, too. But, now consider a rather different case. John has told everyone that he is going to see *The Sound of Music*, which he does indeed see, but then he makes a clandestine visit to another cinema showing a very different sort of film. You believe that his wife has found out about this second visit but that John does not realize that she has discovered his misdemeanour. Your accordingly ask John's wife,

> How do you think John enjoyed the film that he doesn't know that you know he saw last night?

where your intention is again plainly attributive: you are picking out whichever film it was that John does not know that his wife knows that he saw. Unfortunately, you are mistaken: there is no such film. John knows that his wife has found him out, and she knows that he knows. Nevertheless, your attributive designation is perfectly intelligible to John's wife even though there is nothing that fits it. She can readily construct a model of your beliefs that contains a film that her husband does not know that she knows he saw. What makes this step so easy, of course, is that the existence of such a film was precisely her husband's original intention.

The moral of the example, which vitiates other accounts of the referential–attributive distinction (see, for example, Stalnaker 1972; Kaplan 1977; Kripke 1977; Wilson 1978), is that discourse models are constructed on the basis of inferences from general knowledge and a specific knowledge of context. But, it is a mistake to talk of *the* context of an utterance: there is one context for the speaker and another context for the listener. Potential discrepancies between the two models are important both for motivating discourse in the first place and for their effects on a speaker's referential intentions. For example, a speaker may deliberately intend to make it clear that there is such a discrepancy, e.g.:

I don't want to tell you anything about the person I met yesterday.

Here, the definite description is used in a way that contrives to be both referential for the speaker and attributive for the listener. Indeed, Garnham and I argue that the distinction in usage must be made twice over: once for the speaker and once for the listener. It depends on the knowledge that the speaker intends to be relevant to the interpretation of the utterance. In an attributive use, no other unique descriptions that fit the designatum are intended to be relevant to the interpretation of the utterance, even if they are known to the speaker or to the listener. To use a description referentially, a speaker must have a certain minimal knowledge about what it designates, and we suggested that the speaker must have sufficient information to pick it out in at least one other independently individuating way. The underlying principle is that speakers can only refer to those entities on which they can take at least two independent 'cross-bearings'. In a referential usage, they are committed to the substitution *salva veritate* of other of the relevant entity's designations.

The coherence of discourse

Many theorists have struggled with the problem of what makes discourse coherent. Kintsch & van Dijk (1978) argue that there are two levels of discourse structure: microstructure is created by interrelating 'propositions' according to overlaps between their respective arguments, and macrostructure is constructed according to large-scale principles that interrelate the units of text. Several psychologists have independently proposed that there are 'story grammars' that formally specify the large-scale structure of certain stereotyped varieties of story in much the same way that conventional grammars generate the syntactic structure of sentences. One can caricature this idea by proposing the following rule that applies to any story:

STORY → (BEGINNING) (MIDDLE) (END)

When the full rule is used, one has what might be called the Terence Rattigan 'well-made-story' grammar. But many stories start in the middle, others have no middle at all, and some, like Sartre's *Huis Clos*, have no ending. Hence, each constituent is optional. Moreover, to account for 'flashbacks', there is a movement transformation that shifts constituents around. What this

mock example shows, as Garnham *et al.* (1981) have pointed out, is that there are no explicit accounts that identify the membership of such categories as BEGINNING, MIDDLE and END, or other such non-terminal nodes in actual story grammars. Likewise, no arguments have been advanced to justify recourse to the power of context-free rules or transformations. Moreoever, if a story were found that violated the rule above, then a theorist could always claim that it represented a new type of story to which the grammar was not intended to apply.

Obviously, actual story grammars are more plausible than the one I have illustrated. But they rely very heavily on what the users of the grammar bring to it, and they fail to make explicit much that would be required in an algorithm for interpreting stories. Perhaps their main virtue is that they have led to research into the comprehension and memory of stories (Mandler & Johnson 1977; Rumelhart 1975; Thorndyke 1977). Their major empirical claim is that certain stories have a definite structure, which is independent of their content, and which people know and use in the course of comprehension (Mandler & Johnson 1980).

But does discourse really have a macrostructure? One way in which we can find out is by using the old method for generating statistical approximations to English, but with a larger basic unit. The original technique used a series of subjects to generate successive words of a sentence, e.g. for a second-order approximation each subject saw one previous word and added one word of his or her own, and for a fourth-order approximation each subject saw the three most recently produced words and added a fourth. To generate discourse, the same technique can be used, but instead of reading a number of words and then adding one word, the subjects read a number of sentences and then add a sentence of their own. Here is an example of a story that is a second-order approximation (i.e. each subject read one previous sentence and then added a new sentence).

> The baying of the hounds and the screaming of the chickens echoed below me, as I quickly scanned the tracks leading towards the hole – this was going to be a hectic breakfast. I thought I'd better eat a full meal because of the task ahead and the difficulties I might encounter. But it was only when I had cooked myself a steak, and that piece of shark meat that had been ignored by everyone, that I discovered that I could only pick at these titbits, having, as I now recalled, breakfasted, lunched and dined to repletion already. Rather than throw the food away, I rang up my husband at work and asked him to bring home some colleagues to dine with us.

And here is an example of a second-order approximation to a dialogue.

A: If you think I'm going to crucify myself again in order to salvage your self-esteem, it's time for some kind of a fundamental adjustment.
B: I never said you had to crucify yourself just for my sake, but simply that you'd have to make some sacrifices to preserve my sanity.
A: Why should I be responsible for your mental state?
B: I didn't suggest you should be – anyway you're incapable of looking after your own.
A: Well, I wouldn't mind that – a busdriver's not a bad job – just because I've never got a job it doesn't mean I can't look after my own.

The interest of the technique, which was first used in an unpublished study by Kenneth Pease, is not that it generates genuine statistical approximations to discourse – there is not enough real discourse for the idea of a transitional probability from one set of sentences to another sentence

to be anything more than an abstract fiction – but rather that it introduces a 'window' of a variable size through which anaphoric references must operate. With a second-order approximation, anaphora and ellipsis cause problems. In the conversational example, one subject contributes the sentence

B: I didn't suggest you should be – anyway you're incapable of looking after your own.

where it is clear from the previous sentence that what is at issue is B's mental state. The next subject, who obviously has no way of recovering this reference, makes the plausible inference that *your own* refers to A's family, and adds the sentence

A: Well, I wouldn't mind that – a busdriver's not a bad job – just because I've never got a job it doesn't mean I can't look after my own.

The resemblance to real discourse naturally grows stronger in the higher-order approximation. Here is a fourth-order approximation to a narrative.

> What point was there? The dog knew her as a friend and not as an insurgent worth barking over. Anyway the last time they had met, the dog's enthusiastic warning had been rewarded with a kick. She bent down to pat his head and whispered a few reassuring words, then walked over to the door. He lowered his body and shuffled into the corner of the room, and curled up, watching. 'Stupid animal', she thought. His ears followed her down the stairs and into the silence beyond. As soon as he was sure she was gone he jumped up. He would show her who was stupid. No one was going to talk to him like that.

The only thing that goes seriously wrong is the Kafka-like metamorphosis from dog to person. Fourth-order approximations to conversations appear more natural, perhaps because their major referents tend to be the conversationalists themselves.

A: Again – is this the nineteenth breakdown?
B: Don't be so facile – I'm asking you to help.
A: I'm sorry, but I think that being depressed is just a state of mind.
B: It's my mind O.K.? It's in a mess, so help me!
A: I really think you enjoy going through this every so often.
B: But I had thought you were the only person who could understand why I felt like this now.
A: You've become addicted.
B: Now that really is unfair. How can you be so cruel as to term friendship and love just an 'addiction'.

At the level of a sixth-order approximation, the texts seem wholly convincing. The following narrative might be from a Russian novel.

> 'I'm not afraid you know,' Rubashov cried out. The door remained silent and unyielding. It was true, he was next, after Yoshenko. He had heard Yoshenko in the night, crying out. He had listened for the words, but could not make them out, like all the other nights. Had he known it would be his turn next? Of course he did. They all did, secretly, privately. He knew it himself, and like a fly caught in the palm of a hand he was just waiting to be crunched. And there was nothing he could do.

The following conversation, a sixth-order approximation, is commonplace.

A: You still haven't answered my question.
B: I became an adult at seventeen when my father died.
A: But how can you be so precise?
B: First you force me to answer in your terms, then you criticize me for it.
A: I just don't understand your insistence on the importance of change.
B: You don't understand because you think of a change to maturity as a once and for all step.
A: But is change what we ought to focus on?
B: What alternative is there?
A: Surely maturity is when you're aware you're an adult.
B: Doesn't it also involve having all the emotional capabilities and strengths that go with responsibilities of adulthood?

Prose probably has tighter constraints on the sequence of events than conversation. But even prose can vary considerably, as a simple demonstration can establish. The procedure derives from an amusing study also carried out by Pease (1969). He was interested in the extent to which different politicians stuck to the point in answering questions put to them by interviewers. He took transcripts of interviews and cut them up into separate sentences, and then asked a panel of subjects to try to match up the questions with their actual answers. The extent to which the subjects were able to carry out this task provides a measure of the degree to which the politicians stuck to the point. What the experiment showed, of course, was that some politicians were more devious than others. For many years, we have carried out a rather similar study in laboratory classes. Three different sorts of prose passage such as a recipe, a passage from a story, and a fragment of a sociological argument, are cut up into separate sentences. The students' task is to reassemble each passage into its original order. Not surprisingly, it is much easier to reassemble the recipe – though no one has ever got it quite right – than to reassemble the sociological argument. The passage from the story is midway in difficulty. It is plain that the subjects cannot be using a story grammar to help them in this task. Hence, on what basis are they able to reconstruct the sentences? The answer to this question provides us with a general theory about the interpretation of discourse.

A necessary and sufficient condition for discourse to be *coherent* as opposed to a random assemblage of sentences is that it is possible to construct a single mental model from it. Coherence must be distinguished from *plausibility*, since a discourse may be perfectly coherent yet recount a bizarre sequence of events. Coherence depends on one principal factor: co-reference. Plausibility depends on being able to interpret discourse in an appropriate temporal, spatial, causal and intentional framework: a framework that, as Miller & Johnson-Laird (1976) argued, cross-classifies with all semantic fields. When our subjects reconstruct the order of a passage of prose, then they make use of clues about both co-reference and plausibility.

The one substantial hypothesis about the knowledge that underlies plausibility is the notion of a *script* (cf. Minsky 1975; Schank & Abelson 1975). A script represents the normal sequence of events in some relatively stereotyped activity such as dining at a restaurant. Schank & Abelson have written computer programs that represent such scripts and use them to make inferences in understanding stories. The possession of a script allows a speaker to leave many things unsaid with the certainty that the listener will fill in such gaps by default. For example,

it is unnecessary to state explicitly that a customer in a restaurant eats the food that he has ordered. In accordance with the conversational conventions delineated by Grice (1975), it is only necessary to describe such untoward circumstances, as say, the customer refusing to eat his meal – provided that enough has been said to elicit the appropriate script.

The main difficulty with the doctrine of scripts is that knowledge is also used to understand discourse about events that are *not* stereotyped. You can understand Kafka's *Trial* without having a script for the persecution of an individual by an anonymous bureaucracy: the novel is indeed the original 'script' for all such encounters. Likewise, if I tell you a story about an artistocratic detective running a ski-school in Oregon, then I can rely on your knowledge to support many inferences that do not derive from scripts. Schank (1980) is, of course, sensitive to these problems and to the need to account for the acquisition of scripts, but there is still much work to be done to account for the organization and mobilization of knowledge that underlies the plausibility of discourse.

The coherence of discourse depends on its pattern of co-references. Narrative texts depend primarily on a chain of co-references linking one sentence to the next, whereas descriptive texts may depend on references back to the same common topic. One of the main sources of evidence in support of story grammars is that jumbled versions of stories are much harder to remember than the original stories. However, when the order of the sentences in a story is randomized, continuity of reference is also destroyed. Consider, for example, this brief story derived from Rumelhart (1975).

> Jenny was holding on tightly to the string of her beautiful new balloon. She had just won it and was hurrying home to show her sister. Suddenly, the wind caught it and carried it into a tree. The balloon hit a branch and burst. Jenny cried and cried.

When the order of the sentences is randomized, referential continuity is disrupted.

> She had just won it and was hurrying home to show her sister. Suddenly, the wind caught it and carried it into a tree. Jenny was holding on tightly to the string of her beautiful new balloon. Jenny cried and cried. The balloon hit a branch and burst.

It ceases to be clear quite what the story is about: the first sentence refers to *she* and *it*, and the reader is likely to imagine a girl with some sort of prize. Later, when Jenny and her balloon are introduced, it seems that reference is being made to a new person with a new possession. It is simple enough to restore referential continuity by modifying the noun phrases in an appropriate way.

> Jenny had just won a beautiful new balloon and was hurrying home to show her sister. Suddenly, the wind caught it and carried it into a tree. Jenny was holding on tightly to the string of her balloon. She cried and cried. It hit a branch and burst.

Now the text merely reports a slightly implausible sequence of events. Jenny seems to have been holding on to the string of her balloon after it was carried into the tree.

This simple illustration should clarify the distinction between coherence and plausibility. The fact that both are important in the interpretation of discourse has been demonstrated experimentally by Garnham *et al.* (1981). In one experiment, the subjects were given three versions of stories and descriptive passages: the original version, a randomized version, and a randomized version in which referential continuity had been restored. Of course, a subject only

saw one version of any particular story. There was a systematic trend in the memory for the passages. The original stories were remembered better (71% of words recalled) than the random versions (30% of words recalled), but the restoration of referential continuity ameliorated the effects of randomization (43% of words recalled). The descriptive passages, however, had little referential continuity from one sentence to the next – they referred back to the same underlying topic – and, as predicted, the effects of randomization on comprehensibility and memory were negligible. The subjects also rated the comprehensibility of the passages, and their ratings reflected the same pattern.

In another experiment, a group of skilled readers and a group of less skilled readers were selected from a population of children 7–8 years old. The only reliable difference between the two populations was their inferential ability in reading tests. Both groups read a series of short stories presented in the same three versions as used in the previous experiment. As predicted, the ameliorating effects on memory of restoring referential continuity in a randomized story were confined to the group of skilled readers. They have sufficient inferential power to establish co-reference even in the context of a somewhat bizarre sequence of events.

The most direct way in which to disrupt referential continuity is to insert extraneous material between the original identification of an entity and a subsequent reference to it. Ehrlich & Johnson-Laird (1981) have shown that this manipulation has a drastic effect on the memorability of discourse. Their subjects listened to three sentences about the spatial relations between four common objects, e.g.

The knife is in front of the spoon
The spoon is on the left of the glass
The glass is behind the dish

and then attempted to draw a diagram of the corresponding layout by using the names of the objects. On the assumption that the subjects would construct a mental model of the layout as they heard each assertion, the task should be straightforward if the assertions (as in the example) permit the model to be built up continuously. But, if the premises are in a discontinuous order such as:

The glass is behind the dish
The knife is in front of the spoon
The spoon is on the left of the glass

in which the first two assertions refer to no item in common, then the task should be very much harder. In this case, a subject must either construct two mental models and then combine them in the light of the third assertion or else simply represent the premises in a superficial linguistic form until the time comes to make the drawing. The results showed a striking confirmation of the prediction: 69% of the diagrams based on continuous assertions were correct, whereas only 42% of the diagrams based on discontinuous assertions were correct. The crucial factor, however, is the presence of an item in a mental model to which subsequent reference can be made. This thesis is borne out by the relative ease of a third sort of ordering of the premises:

The spoon is on the left of the glass
The glass is behind the dish
The knife is in front of the spoon

in which the third assertion has nothing in common with the second. This lack of a common referent does not matter, because the third premise does refer to the spoon, which should be in the mental model since it was introduced in the first assertion. Indeed, this ordering of the assertions yielded 60% correct diagrams, a proportion that was not reliably different from performance with the continuous descriptions.

It is time to take stock of our current position before proceeding to the final part of the paper. I began with a brief account of model-theoretic semantics and its application to natural language. Next, I introduced the psychological notion of a mental model – a representation of the referents and the relations between them that satisfies the proposition expressed by an assertion – and I discussed the advantages and disadvantages of such a system of representation. The chief advantages are that mental models provide the 'interface' between natural language and the world, and that the semantic procedures required for this relation make possible a whole variety of inferential techniques and heuristics that cannot be used with superficial linguistic representations in mentalese. The disadvantage, of course, is that a single mental model of an assertion builds in more information than is actually expressed by the assertion. A picture may be worth a thousand words, but a proposition is worth an infinity of models. I outlined some of the ways in which people seem to cope with this problem, e.g. they refrain from constructing a model of a radically indeterminate description if it is likely to lead them into trouble. The important point about a mental model, however, is that it can always be modified in the light of subsequent information. Moreover, there may well be hybrid representations that include both model-like elements and language-like elements to represent indeterminacies. (The programming language PLANNER, so notably exploited by Winograd (1972), provides just such a form of representation.) I then considered some phenomena that support the primacy of mental models, including the fact that referential usage is unaffected by questions of existence, and the role of referential continuity in understanding and remembering discourse. My final task is to try to give an account of truth in relation to mental models.

Comprehension and truth

Mental models of discourse account for a number of psychological and linguistic phenomena, and it is natural to wonder how they may relate to model-theoretic semantics and formal definitions of truth (see Johnson-Laird (1979) for a preliminary exploration of this problem). In fact, Kamp (1980) has proposed a most ingenious solution: he employs both a mental representation of discourse and a model that is a complete representation of the world. Since a discourse model represents only a part of what the world would be like if the discourse is true, Kamp formulates the following definition of truth. A text represented in a discourse model is true provided there is a proper embedding of the discourse model in the real world model, i.e. a mapping of the individuals and events in the discourse model onto the individuals and events in the real world model in a way that preserves the same properties and relations.

This approach provides an elegant account of temporal reference. Most previous theorists have assumed an underlying semantics for time consisting of an infinite sequence of durationless moments over which such relations as *later than* can be defined. However, Kamp (1979) proposes that the hearer constructs a discourse model in which tensed verbs and other temporal expressions refer to *events* of finite duration. On the assumption that any pair of events either overlap each other in time or else one event wholly precedes the other, the linear sequence of durationless moments can be reconstructed for the real world model as an idealized limit of

the event structure. Thus, a given event can be treated as both 'punctual' in the discourse model and as extending over a divisible period of time in the real world model: an analysis that appears to be precisely what is called for. It is a common observation that a discourse such as

> He left the room and walked down the corridor. He stopped outside the director's office, lit a cigarette, and walked straight in without knocking.

refers to a sequence of events whose temporal order corresponds to the order of the sentences. This sequence can also be directly represented in the discourse model.

Kamp has introduced a device new to model-theoretic semantics – a discourse model that mediates between language and model-structure – to give a fuller account of the truth conditions of texts. The need for such intermediary models reflects the fact that natural language is both made and used by human beings, and they certainly rely on representations of discourse. There remains, however, the problem of indeterminacy that must be solved in order to relate discourse models to model-structures. Earlier, I discussed a number of alternative strategies for coping with it. Kamp offers another one: information is represented in a discourse model in such a way that it is never necessary to modify it in the light of subsequent assertions in the discourse. He argues that this condition is necessary if a discourse model is to function as a partial description of how the world should be, given that the discourse is true. As he points out:

> The content of an existential sentence has been exhausted once an individual has been established which satisfies the conditions expressed by the indefinite description's common noun phrases and by the remainder of the sentence. But a universal sentence cannot be dealt with in such a once-and-for-all manner. It acts, rather, as a standing instruction: of each individual check whether it satisfies the conditions expressed by the common noun phrase of the universal term; if it does, you may infer that the individual also satisfies the conditions expressed by the remainder of the sentence. This is a message that simply *cannot* be expressed in a form more primitive than the universal sentence itself.

This approach is reminiscent of the treatment of universals in PLANNER (see Winograd 1972).

From a psychological standpoint, there is little doubt that people *do* construct discourse models that they have to revise in the light of subsequent information, and that they do represent universal sentences in a way that is nearer to being once-and-for-all than Kamp's system of representation.

If someone is talking to you about a journey, and remarks,

> There was a fault in the signalling system. The crash led to the deaths of ten passengers.

then you are likely to infer that the passengers were killed in the crash. The speaker has not made this assertion, and might even continue:

> They were arrested when the aeroplane crashed, and subsequently shot as spies.

Plainly, you jumped to the wrong conclusion, though the speaker was partly to blame. Grice's (1975) cooperative principle of conversation might indeed be extended to cover this problem: make your conversational contribution such as is required by the accepted purpose of the exchange in which you are engaged and such that your hearers will not be required to modify their interpretation of it in the light of your subsequent contributions. But, such an idealization is seldom likely to be achieved, and psychologists have no option but to consider an alternative

treatment of discourse models: *a discourse model is a single representative sample from the set of models satisfying the discourse.* This principle requires some elucidation.

First, a discourse model embodies only a limited number of individuals, and, as Kamp has emphasized, it is a partial submodel of the real world model. Secondly, the notion of a *representative sample* does not imply that the set of models satisfying the discourse is constructed and then a sample is selected from them. Comprehension normally leads to the construction of only a single model. The point to be stressed is that the representation of discourse depends both on the model and the procedures for manipulating and evaluating it. These procedures operate according to the principle that any given discourse model is only one of an indefinitely large set of alternatives. The particular model stands in for the set as a whole, and, since it is constructed on the basis of plausible inferences from general knowledge, it is an exemplar of the likely situation described by the discourse. (There is an obvious relation here to the 'representativeness' heuristic of Tversky & Kahneman (1974).) Since the discourse model is treated as a sample, the interpretative system is equipped with recursive procedures for modifying it in the light of subsequent assertions: the system can, in other words, replace one sample by another. This notion was long ago expressed neatly, though somewhat optimistically, by Hume (1896, vol. 1):

> ...after the mind has produced an individual idea, upon which we reason, the attendant custom, revived by the general or abstract term, readily suggests any other individual, if by chance we form any reasoning that agrees not with it. Thus, should we mention the word triangle, and form the idea of a particular equilateral one to correspond to it, and should we afterwards assert, *that the three angles of a triangle are equal to each other*, the other individuals of a scalenum and isosceles, which we overlooked at first, immediately crowd in upon us, and make us perceive the falsehood of this proposition...

In practice, the processing capacity of working memory severely limits the extent to which recursive reconstructions of mental models can be carried out. People make mistakes. Yet, in the absence of a mental logic, the only way in which an inference can be made is by searching for alternative models of the premises that render a putative conclusion false (see Johnson-Laird 1980). And this stratagem is feasible only if subjects construct models of universally quantified assertions by imagining an arbitrary number of individuals with the required properties, a procedure that obviously violates Kamp's principle of not representing any information in a discourse model that is liable to revision.

How, then, is truth to be defined with such a conception of a mental model? In my view, the way to proceed is to take advantage of Kamp's essential insight and to argue that a discourse is true if there is a proper embedding of at least one of its discourse models in the real world model.

Logicians have only related language to models in various ways; psychologists have only related it to the mind. The real task, as I have remarked before (Johnson-Laird 1979), is to show how language relates to the world through the agency of the mind. This is the heart of the problem of comprehension. One active line of research might be termed the 'psychologizing' of model-theoretic semantics: the recognition that many aspects of comprehension are best thought of as constructive processes that yield models of discourse.

There remains one deep philosophical problem: the ontological status of the real world model into which discourse models are embeddable. If linguistic expressions could be mapped onto

objects in the real world as readily as they can be mapped onto objects in model-structures, then one fancies that semantics would not exist. It would hardly be necessary. Unfortunately, the human mind is the only device currently capable of carrying out the mapping. However, it must *construct* its representation of reality, too. If our knowledge of reality is nothing more than another mental model – albeit one that owes its causal origins in part to the nature of the world – then mental models can be rather more than surrogates for reality.

This research was supported by a grant from the Social Science Research Council. I am grateful to many individuals for ideas, help, and criticism, and thank particularly Bruno Bara, Kate Ehrlich, Alan Garnham, Gerald Gazdar, Steve Isard, Hans Kamp, Christopher Longuet-Higgins, Kannan Mani, Jane Oakhill, Ken Pease, Stan Peters, Stuart Sutherland, Patrizia Tabossi and Eric Wanner.

References (Johnson-Laird)

Boolos, G. S. & Jeffrey, R. C. 1974 *Computability and logic.* Cambridge University Press.

Cole, P. 1978 On the origins of referential opacity. In *Syntax and semantics* (ed. P. Cole), vol. 9 (*Pragmatics*). New York: Academic Press.

Collins, A. M. & Quillian, M. R. 1972 Experiments on semantic memory and language comprehension. In *Cognition in learning and memory* (ed. L. W. Gregg). New York: Wiley.

Donnellan, K. 1966 Reference and definite descriptions. *Phil. Rev.* **75**, 281–304.

Ehrlich, K. & Johnson-Laird, P. N. 1981 Spatial descriptions and referential continuity. Preprint, Laboratory of Experimental Psychology, University of Sussex.

Fodor, J. A. 1975 *The language of thought.* New York: Crowell.

Fodor, J. D., Fodor, J. A. & Garrett, M. F. 1975 The psychological unreality of semantic representations. *Ling. Inquiry*, **4**, 515–531.

Frege, G. 1952 On sense and reference. In *Translations from the philosophical writings of Gottlob Frege* (ed. P. T. Geach & M. Black). Oxford: Blackwell. (Originally published in 1892.)

Garnham, A., Oakhill, J. & Johnson-Laird, P. N. 1981 Referential continuity and the coherence of discourse. *Cognition.* (In the press.)

Grice, H. P. 1975 Logic and conversation. In *The logic of grammar* (ed. D. Davidson & G. Harman). Encino, California: Dickenson.

Hintikka, J. 1969 *Models for modalities.* Dordrecht: D. Reidel.

Hume, D. 1896 *A treatise of human nature.* Oxford: Clarendon.

Johnson-Laird, P. N. 1975 Models of deduction. In *Reasoning: representation and process in children and adults* (ed. R. J. Falmagne). Hillsdale, New Jersey: Erlbaum.

Johnson-Laird, P. N. 1978 Mental models of meaning. Paper delivered at the Workshop on Computational Aspects of Linguistic Structure and Discourse Setting, University of Pennsylvania.

Johnson-Laird, P. N. 1979 Formal semantics and the psychology of meaning. Paper delivered at the Symposium on Formal Semantics and Natural Language, University of Texas at Austin.

Johnson-Laird, P. N. 1980 Mental models in cognitive science. *Cogn. Sci.* **4**, 71–115.

Johnson-Laird, P. N. & Garnham, A. 1980 Descriptions and discourse models. *Ling. Phil.* **3**, 371–393.

Kamp, H. 1979 Events, instants, and temporal reference. In *Semantics from different points of view* (ed. R. Bauerle, U. Egli & A. von Stechow). Berlin.

Kamp, H. 1980 A theory of truth and semantic representation. Preprint, Center for Cognitive Science, University of Texas at Austin.

Kaplan, D. 1969 Quantifying in. In *Words and objections: essays on the work of W.V.O. Quine* (ed. D. Davidson & K. J. J. Hintikka). Dordrecht: D. Reidel.

Kaplan, D. 1977 Demonstratives: an essay on the semantics, logic, metaphysics and epistemology of demonstratives and other indexicals. Paper prepared for the Symposium on Demonstratives at the meeting of the Pacific Division of the American Philosophical Association.

Karttunen, L. 1976 Discourse referents. In *Syntax and semantics* (ed. J. D. McCawley), vol. 7 (*Notes from the linguistic underground*). New York: Academic Press.

Kintsch, W. 1974 *The representation of meaning in memory.* Hillsdale, New Jersey: Erlbaum.

Kintsch, W. & van Dijk, T. A. 1978 Towards a model of text comprehension and reproduction. *Psychol. Rev.* **85**, 363–394.

Kripke, S. 1977 Speaker's reference and semantic reference. *Midwest Stud. Phil.* **2**, 255–276.

Mandler, J. M. & Johnson, N. J. 1977 Remembrance of things parsed: story structure and recall. *Cogn. Psychol.* **9**, 111–151.

Mandler, J. M. & Johnson, N. J. 1980 On throwing out the baby with the bathwater: a reply to Black and Wilensky's evaluation of story grammars. *Cogn. Sci.* **4**, 305–312.

Mani, K. & Johnson-Laird, P. N. 1981 The mental representation of spatial descriptions. *Memory Cogn.* (In the press.)

Miller, G. A. 1979 Images and models, similes and metaphors. In *Metaphor and thought* (ed. A. Ortony). Cambridge University Press.

Miller, G. A. & Johnson-Laird, P. N. 1976 *Language and perception.* Cambridge University Press; Harvard University Press.

Minsky, M. 1975 Frame-system theory. In *Theoretical issues in natural language processing* (ed. R. C. Schank & B. L. Nash-Webber). Preprints of a conference at M.I.T., June, 1975. (Reprinted in *Thinking: readings in cognitive science* (ed. P. N. Johnson-Laird & P. C. Wason). Cambridge University Press, 1977.)

Montague, R. 1974 *Formal philosophy.* (ed. R. H. Thomason). New Haven: Yale University Press.

Pease, K. 1969 The great evaders. *New Society*, 2 October, pp. 507–509.

Rumelhart, D. E. 1975 Notes on a schema for stories. In *Representation and understanding* (ed. D. G. Bobrow & A. M. Collins). New York: Academic Press.

Russell, B. A. W. 1905 On denoting. *Mind* **14**, 479–493.

Schank, R. C. 1980 Language and memory. *Cogn. Sci.* 4, 243–284.

Schank, R. C. & Abelson, R. P. 1975 Scripts, plans, and knowledge. In *Int. Joint Conf. Artif. Intell.*, vol. 4.

Smith, E. E., Shoben, E. J. & Rips, L. J. 1974 Comparison processes in semantic memory. *Psychol. Rev.* **81**, 214–241.

Stalnaker, R. C. 1972 Pragmatics. In *Semantics of natural language* (ed. D. Davidson & G. Harman). Dordrecht: D. Reidel.

Stalnaker, R. C. 1978 Assertion. In *Syntax and semantics* (ed. P. Cole), vol. 9 (*Pragmatics*). New York: Academic Press.

Stenning, K. 1977 Articles, quantifiers, and their encoding in textual comprehension. In *Discourse processes: advances in research and theory* (ed. R. O. Freedle), vol. 1 (*Discourse production and comprehension*). Norwood, New Jersey: Ablex.

Strawson, P. F. 1950 On referring. *Mind* **59**, 320–344.

Thorndyke, P. W. 1977 Cognitive structures in comprehension and memory of narrative discourse. *Cogn. Psychol.* **9**, 77–110.

Tversky, A. & Kahneman, D. 1974 Judgement under uncertainty: heuristics and biases. *Science, N.Y.* **185**, 1124–1131.

Webber, B. L. 1978 A formal approach to discourse anaphora. *B.B.N. tech. Rep.* no. 3761. Cambridge, Massachusetts: Bolt, Beranek & Newman, Inc.

Wilson, G. 1978 On definite and indefinite descriptions. *Phil. Rev.* **87**, 48–76.

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