Jerry R. Hobbs

Artificial Intelligence Center

SRI International

333 Ravenswood Ave.

Menlo Park, California 94025

Against Confusion

Jerry R. Hobbs

To an outsider, particularly to someone doing discourse analysis in an artificial intelligence (AI) framework, the recent controversies in literary theory concerning the nature of interpretation are quite puzzling. One camp claims that the interpretation of a text can be anything. The other side claims that there is a single correct interpretation. But all of this confusion can be swept away by a simple observation: in mathematical terminolgy, interpretation is a function of two arguments, the text and a set of beliefs. In interpreting a text, one therefore presents not only an interpretation but also the set of beliefs that warrants the interpretation. One can then go on, if one wishes, to ask the separate question of whether one set of beliefs has a more privileged status than another. Viewed in this light, the controversies are as if one camp said that the mathematical operation of *multiplication* was hopelessly indeterminate because in the context of 2 the product of 2 is 4 whereas in the context of 5 the product of 2 is 10, while the other camp claimed that, no, the product of 2 is always 4.

AI provides us with a technical vocabulary that makes it possible to be somewhat more precise and detailed than is customary in discussing processes of interpretation. In Part II of this article, I present a framework, along with a corresponding technical vocabulary, that has proved useful in investigating discourse interpretation from an AI perspective. Among other things, it allows us to explicate the roles of intention and belief in interpretation. We will then be in a position to examine several characteristic views in literary theory in terms of the framework.

There has been a recent and widely discussed claim that it is incoherent to separate meaning and intention. Since this distinction is crucial in what I present, I begin in Part I by responding to this claim.

I

Steven Knapp and Walter Benn Michaels ("Against Theory") have argued, or rather asserted, that meaning is an incoherent notion in the absence of an author's intention. It is certainly true that in the canonical case a text has an author who intends to convey something, and that something is what we call the meaning. "What did you mean?" and "What were you intending to say?" are often taken as equivalent. To reinforce this identification, Knapp and Michaels have us imagine that as we are walking along the beach, we see a wave wash up and, receding, reveal a poem written in the sand. We will believe either that the poem was written by some spirit of the sea capable of intentions, or that the marks in the sand resulted from some hugely improbable coincidence. Knapp and Michaels have the following intuition: "... in the second case—where the marks now seem to be accidents—will they still seem to be words? Clearly not. They will merely seem to *resemble* words" ("Against Theory" 728). The marks have no author, are thus not language, and thus have no meaning.

This is the whole "argument." Unfortunately, I have the opposite intuition. It seems to me that the marks in the sand *are* words and *do* mean something. The event would not be remarkable otherwise. In any case, neither their intuition nor mine is worth very much, both being theory-laden. The example is so implausible it is doubtful whether anybody could have very firm intuitions about it. Let us consider three more commonplace examples to see if it is possible for texts to mean something independent of an author's intention. Here I will be appealing to the reader's everyday intuitions about the word "intention"; in Part II, I work toward a more precise analysis of the term.

The first example is printer's errors. A favorite of mine appeared in a New York Times article on the voyage of the Pioneer 10 spacecraft beyond the solar system. Toward the end of the article, the writer intended to say, "Pioneer 10 carries a message ... in the form of a plaque designed to show ... the place and time where it began its long journey." Instead the newspaper printed, "Pioneer 10 carries a message ... in the form of a plague designed to show ... the place and time where it began its long journey." Let us suppose this was indeed a printer's error and not sabotage. The fact that what was printed does not correspond to any author's intention in no way diminishes our enjoyment of it, and it is hard to see how we could enjoy it if we did not first interpret it, that is, determine what it means. This means something, and it means something *other* than what the author of the article intended.

The next example comes from the world of computers. Before giving the example, I will give three negative examples for purposes of orientation. I log onto my terminal in the morning, and on the screen I see the text, "Good Morning!" It is a text and it has meaning, but I do not need to attribute intention to the computer or deny that an intention lies behind the text. The programmer, whoever and whenever, was the author, and the text means what he or she intended it to mean.

Next consider a computer program that generates random sequences of English words. We look over the output of the program and find some sequences that approach genuine poetry. There is too much distance between the program and the output for us to call the programmer the author. The words might have been read in from a file the programmer never looked at, and the random-number generator might have been a library subroutine whose code the programmer never inspected. But it is quite reasonable to say that the sequence of words is not a text at all, but simply an object on which we have chosen to impose some interpretation, as a kind of play, in much the same way as we might see the shape of a dog in a cloud. We certainly would not act on the content of the text. If we found the words "Ronald Reagan is a communist," for instance, we would not thereby come to believe that Ronald Reagan is a communist.

Next consider a program that "plans" its utterances, of the sort that has been implemented by AI researchers. It has a goal, that is, a logical formula or other data structure representing the condition to be achieved. It has knowledge, again in the form of logical formulas or other data structures, about what kinds of states or actions cause or enable what other kinds of states and actions. There is a process, called "planning," that uses this knowledge to decompose the goal into subgoals and these into further subgoals, until it derives a sequence of executable actions—in this case, the utterance. Again, there is too great a distance between the program and its output for us to call the programmer the author of the output. If the system is in practical use, say, telling us how to find something or how to use or fix an appliance, we had better take the utterance as a meaningful text and act on its meaning. But a reasonable case can be made (although many balk at this) that the program itself has intentions. If we want to be especially concrete, we can say the goals and subgoals are its intentions. The whole structure of the program is informed by the folk psychological theory and vocabulary of intentional action, making attribution of intention quite natural. In this case, the text has meaning, and it means what the program intended it to mean.

But let us now consider an example that falls in the middle of these three cases. We do not have to search far. A pocket calculator will do. Suppose I type in "1129.35 – 959.47", and the calculator responds with the text "169.88". I'm certainly not the author of the text; I might even be surprised at what I see. Neither the designer nor the manufacturer of the calculator could be called the author; the distance is too great; it is extremely improbable that either of them ever considered my particular subtraction problem. The sequence of numbers is a meaningful text; I interpret it using the same rules of interpretation I would use if a human had typed it out in response to my question. My interpreting it is not merely playful activity; I might enter it on my income tax return, sign my name at the bottom, and become legally responsible for my interpretation. Finally, we would not want to attribute intention to a pocket calculator. To do so would be a trivialization of the notion of intention and a consequent trivialization of the point Knapp and Michaels had hoped to make. The text "169.88" is a meaningful text with no human author and no intention behind it.

A final example that drives a wedge between meaning and author's intention is provided by Japanese linked poetry. In a group of three or four poets, one composes a stanza. Another poet makes up a second stanza related to the first in some way. A third poet composes a third stanza that is related somehow to the second but not necessarily to the first. The poets continue to alternate in this fashion for 36 stanzas, to produce a poem that goes through quite a number of twists and turns. It is quite common for a new stanza to force a reinterpretation of the preceding stanza, changing the implied locale, the circumstances, the sex and condition of the agent, and even the meanings of words. Very often, one suspects, the reinterpretation would surprise and delight the preceding stanza's author. A typical stanza thus has two meanings, one corresponding to its author's intention and determined by its link to the preceding stanza, and one constructed by the author of the following stanza and determined by its link to that stanza. Moreover, both meanings are essential to the working of the complete linked poem.

All three of these kinds of text are intentionless (or, in the third case, doubly intentioned), but they are hardly "accidental likenesses of language," and they have meaning. Though commonplace, they are admittedly marginal, but like many marginal phenomena they allow us to see clearly distinctions that are blurred or masked in more central cases. They show that meaning and author's intention do not coincide.

There is another (uninvited) possible reading of Knapp and Michaels' article. They could be saying that to interpret something as a text, we must imagine an agent's intention as its source. The temptation of this position is clear. Since we are so adept at reasoning about human action, it often helps to imagine people in control where there aren't any.[1] But this is hardly necessary; the above examples show that we have ample experience with intentionless texts. So Knapp and Michaels, under this reading, could only be *stipulating* a new meaning for "intentional"; it is synonymous with "interpretable." Their argument then reduces to the following trivial one. We stipulate "x is intentional" to be equivalent to "x can be interpreted." Therefore, to be interpreted, an entity must be intentional.[2] The effect of this stipulation is to make the word "intentional" unavailable as a technical term; "interpretable" will suffice. But in the AI framework explicated below, "intentional" and "interpretable" both turn out to be useful technical terms, and their meanings differ.

II

There is a technological aspect to AI—the effort to build smart computer programs—and a theoretical aspect. In the latter aspect, which is the one of interest here, one tries to discover general principles governing intelligent agents, regardless of how the agents happen to be embodied physically. This endeavor proceeds by means of a radical simplification. A computer program, or robot, or "cognitive agent," is constructed, or just imagined, to simulate, or duplicate, some intelligent behavior humans are capable of. This behavior is modeled in terms of formal symbol manipulation procedures. Questions about human capabilities, which are tangles of complex interactions and for which we have an inadequately precise vocabulary, are translated into questions about the workings of the cognitive agents, for which we do have a precise, computational vocabulary and where we know, at least in principle, everything that is going on. This translation can isolate the core of an issue, suggest further lines of analysis, and frequently expose the falsity or tautologous character of an argument. We can often get crisp answers to mushy questions. Whether the crisp computational stories we tell about the cognitive agents project back to the human level is never certain. But if one is to argue that the crisp answers do not project back, one must say precisely how humans differ from the cognitive agents in a way that would make the projection fail. In any case, there is a long history in science of successful use of such idealizations.

The radical simplification is this. A cognitive agent possesses a set of beliefs. In AI this is generally called a "knowledge base" since one typically wants one's robot to believe true things. But because we will also want to include false and uncertain beliefs, opinions, values, heuristic strategies, and so on, we will call it a "belief system." One useful way of viewing a belief system is simply as a set of logical formulas encoding the agent's beliefs about the physical and social world in which it finds itself. The belief system includes not just general knowledge, but also a model of the immediate situation or environment—a theory of what is going on right now, including expectations, or beliefs about future events. The agent is linked to the world by means of various sensors and effectors. Its beliefs must be in accord with what it senses, and it will act in accord with its beliefs.

Next we can imagine a society of such agents, each with its own belief system. Suppose they can communicate, that is, produce and receive utterances via some medium. Then each agent's belief system must include beliefs as to what other agents in the environment believe and what beliefs it shares with them. Thus, beliefs must have more than just their content encoded; they also need to be tagged with information about who else believes them and, in particular, about what groups mutually believe them.[3] Conventions may be represented in this fashion, including the conventions of language.

For purely computational reasons, we may assume that some particular subset of beliefs is active or in focus at any given moment. Only these beliefs are used by the agent's internal processes, although the agent also has means of moving beliefs into and out of focus. Alternatively, beliefs may have degrees of focus, where degree of focus determines the order of access to the beliefs by various processes.

The standard view in AI of the agent's procedure for generating utterances and other actions is that it is some sort of planning mechanism, as described above. The agent starts with a goal (an intention) and develops, or begins to develop, a plan of action, that is, a decomposition of the goal into subgoals, and these into further subgoals, ultimately yielding a sequence of actions, such as utterances, which it is believed will achieve the goal. As the actions are executed, the environment is monitored, and when unanticipated conditions arise, the plan is modified to accommodate them. Since utterances are typically intended to affect the beliefs of others, the planning mechanism, in designing the utterance, must take into account the beliefs of the other agents participating in the discourse, and especially those things that are mutually believed. Moreover, it must take into account the interpretation procedures that will be used by the other agents. What is presupposed by an utterance should be mutually known to the others or easily reconstructed by them. In particular, most of an utterance will depend on the conventional meanings of words and an implicit conventional theory about how utterances are understood. The less personal knowledge the participants have about one another, the greater the reliance that must be placed on conventions shared by a larger society to which they all belong.[4]

Thus, for the bare notion of intention, one substitutes a hierarchy of goals and a fairly complex planning and monitoring mechanism, enabling a much more fine-grained analysis of what individual features of texts and other behavior are there to achieve.

AI work in discourse interpretation is characterized by a concern for specifying, with computational precision, how the hearer makes use of his or her commonsense knowledge of the world and the immediate situation to interpret utterances, and in particular how utterances can be related to the speaker's presumed plan. Various accounts have been developed. In what follows I will, unsurprisingly, present my own.

We will assume the agent's interpretation procedure works by translating the utterances (the text), produced by another agent we may call the *author*, into logical formulas and then drawing inferences from its belief system in such a way as to satisfy a set of requirements that specifies just what a "good" interpretation is.[5] What these requirements are is, as they say, a research question, but four very strong candidates are the following. (1) Utterances are anchored referentially in the mutual beliefs of speaker and hearer, and reach out into the speaker's private beliefs in a bid to make new information mutually believed. This referential anchor must be identified and the new information must be recognized as such. (2) Words that are functionally related syntactically should be seen as congruent semantically. This constraint forces the interpretation of many instances of metaphor and metonymy. In the case of metonymy, an explicitly mentioned entity must be "coerced" into an implicit entity that satisfies the constraint. In the case of metaphor, certain inferences about an entity must be assumed or suspended to satisfy the constraint. In

America believes in democracy,

"believe" requires its subject to be a person, so "America" must be interpreted metonymically as standing for something like "the people of America," or it must be interpreted metaphorically, acquiring for the occasion the relevant properties of persons. (3) Different segments of the text should be seen as coherently related, in a way that gives the whole text a unitary structure; this requirement for coherence in texts probably derives ultimately from principles of cognitive economy that people apply and that the agents should apply in attempting to make sense out of the world in general, principles involving things like causal linkage and assumptions that apparently distinct entities are identical. All of these constraints are sometimes violated, but where they are, the violation should be recognized; much of the delight that one derives from violations in literary works comes from our efforts to find a way in which the constraints are *not* in fact violated, to discover some hidden coherence. (4) The text needs to be related to the agent's theory of what is going on in the environment. Typically, but not always, this includes the agent's beliefs about the author's intention, or more generally, the author's plan as it unfolds in time; the agent should try to relate the text to what the agent believes the author is trying to accomplish.

This fourth point deserves expansion, since it is where interpretation and author's intention meet. The first thing to note is the phrase "what *the agent believes* the author is trying to accomplish." In the ideal case, the agent is entirely correct about the author's plan and cares about the utterance's relation to it. But like it or not, the agent, for all it knows, could be a brain in a vat, entirely deceived about what is going on around it. A real robot, especially during debugging, is often deceived in just this way, as its programmers manipulate its sensory inputs to test it. The agent can form good hypotheses about an author's intentions, just as it can about anything else in its environment. But it can never be certain about any of its hypotheses. The most it could hope for is a consistent theory of the author's "psychological" life that will account for all the author's actions it perceives. So the author's intention plays at best an indirect role in the interpretation process: it plays a causal role in some observable actions, which the agent can then use, along with background knowledge, to form a belief about the author's intention. Only this belief can play a direct role in interpretation.

Moreover, among us humans there are many situations in which the author's or speaker's plan is of little interest to the reader or hearer, and we would expect the same to be true for our cognitive agents. Someone in a group conversation may use a speaker's utterance solely as an excuse for a joke, or as a means of introducing a topic *he* or *she* wants to talk about. Very often two speakers in a discussion will try to understand each other's utterances in terms of their own frameworks, rather than attempt to acquire each other's framework. A medical patient, for example, may describe symptoms according to some narrative scheme, while the doctor tries to map the details into a diagnostic framework. A spy learning a crucial technical detail from the offhand remark of a low-level technician doesn't care about the speaker's intention in making the utterance, but only about how the information fits into his own prior global picture. A historian examining a document often adopts a similar stance. In all these cases, the hearer has his or her own set of interests, unrelated to the speaker's plan, and Requirement (4) involves no more than relating the utterance to those interests. In the conversations I have analyzed (see, for example, Hobbs and Evans), I have found this to be the case astonishingly often. Thus, not only is the role of the author's or speaker's intention indirect; it is frequently not very important.

The agent's interpretation procedure works by drawing inferences from its belief system, but two caveats are in order. First, inferences are drawn in a selective fashion, determined by what will lead to a good interpretation. Instances of metaphor and irony are only the most obvious cases in which this control over inference is required. Second, the agent must often assume things to be mutually believed, for no other reason than that it will lead to a good interpretation of the text. David Lewis has called this process "accommodation," and we may call the proposition that is assumed an "implicature," since it is consistent with what H. P. Grice calls "conversational implicature."

We can summarize all of this in a single formula that is applicable beyond the details of this particular theory:

$$F(K, T) = I$$

An interpretation procedure F takes a knowledge base or belief system K and a text T, and produces an interpretation I. Each of these four elements requires some comment. In my comments, I will cease being fastidious about the distinctions between these agents and real people.

T: In general, there should be little dispute about T. Sometimes in conversation, one is not quite sure whether a nonverbal gesture is part of the text or just accidental, and in medieval manuscripts the words are often in doubt. But, for the most part, we can assume that the sequence of words that comprises the text is given.

Someone not familiar with recent literary theory might think this is all there is to say about T. But, as Stanley Fish has pointed out, interpretation goes all the way down. It is not a brute fact that a mark on paper is an instance of the letter "g", but is rather the result of interpretation. There have been, in fact, researchers in pattern recognition trying to make explicit the set of beliefs or interpretation rules that allows us to interpret an arrangement of lines and curves, or at an even lower level, an arrangement of pixels, as the letter "g".

Ultimately, in text interpretation as in every scientific or critical enterprise, we must bottom out in conventionally agreed-upon "evidence."[6] For text interpretation, this first involves a decision or an agreement *that* some physical object exists or *that* some physical phenomenon has occurred. This should not pose any problems. I doubt that any literary critic, as a critic, could seriously maintain that copies of *Ulysses* do not exist as physical objects, regardless of what one may take them to be. If we cannot accept the reality of trees, chairs, and books, it is hard to see why we should care about the feelings of Stephen Daedelus toward Leopold Bloom.

Next there has to be some conventionally agreed-upon account of how the physical entity presents itself to us. This does not seem problematic either, since one can express the account at as low a level as one pleases—for example, in terms of the impingement of light rays on the retina. Disciplines are defined by what they consider given and what they take to be problematic. Generally a literary critic will not be interested in interpretation processes below the level of the word or the letter. It would be acceptable to him or her to take as a fact that the first word of this sentence is "It". One can imagine circumstances, of course, in which it is crucial to determine whether a letter written in pencil is a "g" or a "q", and a microscopic examination may be required. Here the conventionally agreed-upon "facts" will be statements about the depth of the impression, the presence of bits of graphite, and so on.

Finally, one has to decide that this physical entity is to be interpreted as a text. This decision is part of a larger effort to construct the simplest theory, covering the most details, of all the entities one encounters; for some entities, the most economical theory is that they are texts. There are problematic cases, of course; an archaeologist has to decide whether scratches on a rock were carved by people or by a geological process. But the overwhelming majority of the things we decide to call texts give scant support for any alternative treatment. Once these assumptions are made, we are in the game defined by the above formula, and all of the following arguments apply.

Hence, we will assume that the text exists as a physical object, that there is a conventionally agreed-upon set of "facts" about what the object is at some level—whether pixels, letters, or words—and that a decision has been taken to regard it as a text and to apply interpretation rules to it. That is, we can take T to be given.

F: Some indications were offered above as to what the interpretation process looks like. AI researchers in discourse analysis have gone into greater detail in numerous articles. It remains a big problem, but it is a healthy area of research. For the purposes of this article, we will assume the problem is solvable and ask what the consequences are. That is, we will assume F to be given.

It is important to note that there is a trade-off between F and K, between the interpretation process and the belief system used in interpreting. Any particular interpretive principle, such as "In Japanese poetry, the mention of cherry blossoms means that the season is spring," can be viewed as part of the interpretation process—as something we *do* when we interpret—or it can be viewed as a belief that is accessed by the interpretation process—as something we *use* when we interpret. There is no fact of the matter; we can choose either option. For the purposes of this article, we will choose the latter; interpretive principles are beliefs. Individual differences can also be accommodated in this way. It is quite possible that different people have different interpretation procedures, that they use radically different means to comprehend language. But even if this is true, then insofar as we are able to describe the interpretation procedures explicitly, we can factor out the differences, call them differences in belief, and let F be whatever is common to all interpretation procedures. Thus, F need not be indexed by *who* is doing the interpreting or *how* they choose to do it on a specific occasion.

Finally, one might ask why F is a function in the mathematical sense of yielding only one result.[7] Is it plausible to say that F applied to a single text and a single belief system will always yield a single interpretation? What about ambiguity? A purely formal way around this problem is to say that I can be not just a single interpretation, but a set of interpretations. But I think a more satisfactory answer is possible. Generally, when we entertain different readings of an ambiguous text, we do so by shifting something in the belief system we are interpreting the text against. For example, when Hirsch sets out to convince his reader of the pantheistic interpretation of Wordsworth's poem, "A Slumber Did My Spirit Seal," he does so by spelling out Wordsworth's beliefs about "the immortal life of nature." In poems where we are given a few bare details that we can expand into a complete picture in several different ways, we can see our expansions as resulting from different implicatures, that is, different "beliefs" that we assume to be in K in order to accommodate the author.

I: An interpretation I is some formal representation of the content of the text that satisfies at least the four requirements for a good interpretation discussed above. It encodes the information conveyed by the text, the relevant inferences, and implicit structural relations that have been discovered among various elements. For most noncomputational purposes, a rough description in prose of the less obvious aspects will do.

There is, of course, more that one could say about a text than just what is contained in I. We can ask what someone would have to be like to produce such a text. We can ask what function the text performs in the larger social world. As I understand Hirsch, these are questions about the "outer horizon" of the text. I is what I understand by his notion of "inner horizon."

K: The belief system K is intended to include the whole range of beliefs, from simple facts about the physical world to interpretive conventions for particular genres. Interpretation depends on context, and it is in K that the context is encoded. For different authors and different occasions, the agent will have different beliefs about the author's intentions, about what portions of the belief system are shared with the author, and about the current situation. In addition, on different occasions different beliefs will be in focus and different interpretations can result.

It has often been argued that context is unbounded, and that therefore it is impossible to formalize it. (Mailloux has a recent and eloquent statement of this position.) Our knowledge is certainly unbounded in the sense that indefinitely many propositions can be deduced from it, but this is hardly an argument against formalizability; deduction is well understood. The argument must therefore be that there are indefinitely many things one can say about a context beyond what can be deduced. It seems obvious to me, and I think most other AI researchers, that since we are finite creatures with finite access to our environments and a very finite amount of time, there is only a finite amount of context that can be relevant to the interpretation of any situation. In fact, several large-scale efforts are under way to encode the knowledge an agent would need to understand everyday situations, and other projects are directed toward devising procedures for extracting from this knowledge just the parts that are relevant to any particular situation. The formalization of context is still an unsolved problem, but it is a vigorous area of research.

The belief system used in interpretation need not consist only of statements that are actually believed. A statement may also be embedded within a hypothetical context. This is required for understanding fictional texts and texts from other cultures and previous periods of our own, and also for understanding indirect proofs and other counterfactuals. The hypothetical statements enter into the interpretation procedure in exactly the same way as real beliefs, differing only in that they need not accord with what the agent perceives and in that the agent is less likely to act on them. We can flip among these hypothetical contexts with some facility, one time pretending we believe one thing, and another time something else. This is an important point for both discourse analysis and literary theory. Even though we often do not care about the speaker's beliefs in interpreting an utterance, at least as often we do care. In these cases, we can interpret the utterance not with respect to our own beliefs but with respect to our best guess of the speaker's beliefs. Insofar as we read literary works as a way of having conversations with the great minds of the past, it seems reasonable to interpret their texts with respect to *their own* belief systems, to the extent that we can surmise them. In brief, the beliefs used in interpretation do not have to be actually believed. We are not, as some writers try to cast us, prisoners of our own beliefs. We are prisoners of what we can imagine someone believing, and this gives us a much more comfortable cell.

Finally, there is no need to tie K to a real person. The belief system does not have to be *someone's* belief system. In this framework, it is merely the specification of a set of propositions. It can therefore be viewed as standing for the belief system of an ideal reader or an idealized author. It can be an author's real beliefs, or the beliefs he believes he shares with his audience, or the beliefs he wants his audience to think he has. It can be the set of beliefs a reader *should* bring to the text, whether or not anyone ever really does. It can be the set of beliefs that defines some "interpretive community." We can construct idealized, consensual belief systems against which to interpret texts of multiple or indistinct authorship, such as the Constitution or the Bible. By allowing such disembodied belief systems, we can abstract away from irrelevant vagaries of individual readers and writers.

Just what belief system should be used in interpreting a particular text depends on the purposes to which the text and its interpretation are to be put. In particular, what belief systems should be used in interpreting literary texts depends on the function of literary texts in our society. That issue is beyond the scope of this article.

To summarize, then, we may assume that, in the equation, F and T are given and we must determine K and I. We have one equation in two unknowns. This of course does not determine either the belief system K or the interpretation I of the text, but it does place constraints on the possible K-I pairs. We cannot determine a belief system appropriate to the text simply by examining the text. We need to assume a particular interpretation of the text. Similarly, we cannot look at a text and determine its interpretation without making certain assumptions about the underlying belief system. When we understand or analyze discourse, we do so by hypothesizing a K-I pair. We assume an interpretation of the text and a portion of the underlying belief system that will support that interpretation. We can call this pair a "theory of the text." The equation expresses the fact that there are constraints on the possible K-I pairs, the possible theories of the text.

Consider an example. When I first read the opening line of Shakespeare's 68th sonnet,

> Thus is his cheek the map of days outworn,

I had a very powerful image of an old man whose face was deeply wrinkled. These wrinkles were like the roads on the map of the life he had led. Later I read the footnotes. "Map" meant "symbol." "Days outworn" meant "ancient or classical times." The line meant that his face was the symbol of classical beauty—almost the precise opposite of my interpretation. I had interpreted the line against a belief system that included knowledge of Rand-McNally road maps and beliefs about the romanticization of old age. The function of footnotes is to tell the modern reader something of the belief system Shakespeare must have assumed he shared with his Elizabethan reader.

Another example comes from work that the anthropologist Michael Agar and I have done on some life history interviews of a heroin addict. (See, for example, Agar and Hobbs.) He is telling a story, and at one point he says,

> Time was passing.
> I was feeling worse all the time.

For most of us, there is no especially strong relation between these two utterances. But for the addict these sentences are elaborations on the same theme. If we are going to recognize this, we need to assume that very salient in his belief system is the fact that the passage of time implies that junk is running out and he is in need of another fix.

In specifying the details of K, different degrees of formality and precision are required for different purposes. At one extreme, about a decade ago I wrote a long and unreadable technical report giving an excruciating blow-by-blow account of what an interpretation procedure would do with one paragraph from *Newsweek*. The specification of the underlying knowledge base took 43 pages, and the account of what the interpretation procedure did with the text and the knowledge base ran to 58 pages. When one is talking not to computers but to people, as one does in discourse analysis and literary criticism, one can focus on the difficult passages and state only the less obvious beliefs, as I did in the Shakespeare and the junkie examples.

There are many possible theories of a text within the constraints set by the equation. To decide among competing theories, or competing K-I pairs, we try to find the best K and the best I. I have already discussed some of the factors involved in determining how good an interpretation is. The junkie text provides an example. If we can discover the elaborative relation between "Time was passing" and "I was feeling worse all the time," the interpretation has greater structural coherence and is thus better than one that treats the two sentences as unrelated. There are various criteria that determine the appropriateness of a K. For literary texts one often wants the belief system that the author assumed he or she shared with the audience. Theorists who argue for the primacy of the author's intended meaning can be seen as arguing for the use of this belief system. One hypothesis about the belief system is then better than another to the extent that it generalizes over a larger number of texts by the same author or authors from the same culture. The Shakespeare example illustrates this point; the footnotes tell how Shakespeare and other Elizabethans used the words.

A text can be interpreted in many ways. Stanley Fish is adroit at showing how an initially outlandish interpretation can be made plausible, and this might be taken as an argument that a text can mean anything, or that an "interpretive community" can make a text mean anything. But this does not follow. To see how absurd this position is, let us consider what would be involved in constructing a "belief system"—in this case simply a lexicon—that would enable us to read *Paradise Lost* as *Hamlet*. "Of" would have to mean "who's." "Man's" would have to mean "there." (We'll ignore punctuation.) "First" would have to mean "nay." "Disobedience" would have to mean "answer." "And" would have to mean "me." "The" would have to mean "stand." "Fruit" would have to mean "and." But now we encounter a problem. "Of" would have to mean "unfold," but we've already said that "of" means "who's." We can get out of this by having context-dependent rules: following "fruit," "of" means "unfold." It is obvious that our difficulties become compounded the farther we go, and that as we approach the end, each rule for interpretation would be nearly as long as *Paradise Lost* itself. The point of this rather silly exercise is to demonstrate that the set of possible interpretations, large as it is, is really quite insignificant compared with the vast set of impossible interpretations. The requirement that a belief system must be constructed is really quite constraining, given the most rudimentary constraints on the content of the belief system. It means that the space of possible interpretations is onedimensional rather than two-dimensional. We have one degree of freedom, but we do not have two. The difference is precisely the difference between having to stay on the highway and being able to drive all over the landscape.

It is important to emphasize that none of this unduly shackles the discourse analyst or literary critic. There is still plenty of room for his or her unique insights. As in any science, there are no constraints placed on the process of *arriving at* a theory. The constraints are applied in its *validation*. The analyst or critic can appeal to the full range of his or her knowledge of the author's culture and can use unconstrained ingenuity in constructing theories of a text. However, when it comes to validating a theory of a text or deciding among competing theories, he or she must convince us that the hypothesized belief system is appropriate and indeed supports the proposed interpretation. So for validity in interpretation, we do not need the author, as E. D. Hirsch argues; we only need to be explicit about the contributions of the belief system and of the text. All of this is not so different from standard practice. Even Stanley Fish, when he argues for the plausibility of an "Eskimo" reading for Faulkner's "A Rose for Emily" (346), does so by having us imagine that in Faulkner's belief system there is a belief that he is an Eskimo changeling.

Let us briefly examine several popular positions in literary theory in light of this framework. The New Criticism, and Wimsatt and Beardsley's position in particular, can be viewed as an attempt to standardize the belief system. The privileged belief system is an ideal one that includes only those beliefs or facts that an informed, but not too informed, reader would possess. It should include the conventions of language and presumably the facts about the world that are accessible to everyone, such as the fact that stones are not alive, but it should not include facts "about how or why the poet wrote the poem—to what lady, while sitting on what lawn, or at the death of what friend or brother." (Beardsley and Wimsatt 10). Of course, since there is such great divergence among various people's belief systems, one might ask whether the ideal is possible to achieve. For example, should it contain detailed knowledge of the *Odyssey*?

Generally, an author has a specific meaning to communicate to the audience. He or she has beliefs about what beliefs are shared with the audience, and so constructs the text upon this set of beliefs. E. D. Hirsch can be understood as saying that for literary texts the reader's task is to discover this belief system and to interpret the text with respect to it. There are many good arguments for granting this belief system a privileged status. An argument that is not good, however, is that only thus does a text acquire a determinate meaning. It already has a determinate meaning—determined by K and T both. Fix K any way you please, and the meaning is determined by T alone.

Knapp and Michaels, in their sequel to "Against Theory," "Against Theory 2: Hermeneutics and Deconstruction," characterize the hermeneutic position as one that posits a "verbal meaning" of a text which determines its identity but nevertheless allows it to be construed in various ways. Those adopting this position are seeking to explain how the same text can take on different meanings for different readers and different ages. Knapp and

Michaels contend that it is arbitrary to choose verbal meaning as the criterion for textual identity, rather than, say, letters, or verbal meaning plus some bizarre additional rules of interpretation, and that the only coherent notion of meaning is the author's intended meaning. From the perspective of our framework, Knapp and Michaels are correct in saying that verbal meaning is an arbitrary choice—a text can be interpreted with respect to any K. The physical object, or rather the way it impinges upon our senses, is ultimately the only determinant of textual identity, and one can attempt to interpret it with respect to any K at all. The hermeneutic position is correct, or nearly correct, in that it isolates verbal meaning as the choice of K most appropriate for explaining the force of literary texts on readers through the ages. In effect, one partitions K into beliefs of interest and beliefs too low-level to be of interest. One interprets the physical object with respect to the latter set of beliefs and any two objects that yield the same interpretation—two copies of *Ulysses*, for instance—are for the purposes at hand viewed as identical. One then interprets this with respect to the beliefs of interest—verbal meanings, or the conventional meanings of words. The beliefs of interest may coincide with the author's beliefs, in which case the interpretation will be what the author intended, or they may reflect the time and situation of the reader, in which case the interpretation may be quite different from anything the author ever imagined. In any case, Knapp and Michaels are simply wrong in saying that the author's intended meaning is the only coherent criterion for textual identity and the only coherent notion of meaning.

Stanley Fish in the introduction to *Is There a Text in This Class?* says, "In 1970 I was asking the question, 'Is the reader or the text the source of meaning?'" (1). Within the framework we have developed, this is like asking of multiplication whether the multiplier or the multiplicand is the source of the product. The meaning or the interpretation I is a function of both the text T and the reader, parameterized as K. When Fish makes the provocative statement that there is no text until the reader writes it, he is really making the rather more mundane observation that there is more to K and less to T than one might have thought.[8]

The "facts" about the text are constructed, conventional facts, but that is not to say they are arbitrary. There are many "facts" that simply cannot be constructed. The "fact" that aspirin is a painkiller may be a constructed "fact," but it is not a possible constructed "fact" that LSD is a sleeping pill or that the Golden Gate Bridge collapsed in 1984. Our constructions, including our interpretations, are heavily constrained by the way the (not directly accessible) world is. There is no convention-free way to talk about the world, but that does not mean that there is nothing but convention. The world is still there to respond to our actions in ways beyond our control and to enforce a degree of mutual consistency with other agents. The world is experienced primarily (if not entirely) in the constraints it places on the interpretations we construct. The text exists as part of the world and is experienced as a set of constraints on what we can take the text to mean.

In 1979 Fish wrote that "meanings are the property neither of fixed and stable texts nor of free and independent readers, but of interpretive communities that are responsible both for the shape of a reader's activities and for the texts those activities produce" (322). This is an example, common in Fish's writings, of falsely posing several factors as mutually exclusive alternatives, rather than using the list of factors as a starting point in a detailed analysis aimed at discovering the contributions of each. It was stated above that a belief system contains not only the beliefs of the agent, but also an indication of who else holds those beliefs. For each fact P, it contains not just the fact P, but the fact $mutually\text{-}believe(S, P)$, where S is the set of people or agents among whom P is mutually believed. Fish's "interpretive community" is such an S. For an "interpretive community" S to be the source of an interpretation would be for the belief system upon which the interpretation is based to consist entirely of beliefs P for which $mutually\text{-}believe(S, P)$ is also believed. But it is obvious that there is seldom a single such S. Each reader belongs not to one but to a unique blend of many "interpretive communities." A variety of "interpretive communities," cultures, social organizations, shared and private experiences, and original ideas is responsible for a reader's belief system's being what it is, and thus they all contribute indirectly to the reader's interpretations. But it is only the belief system the reader uses that is directly responsible for the interpretations. By making the set of beliefs explicit, including the "interpretive community" associated with each of the beliefs, we can begin to tease out the contributions made by several "interpretive communities" to a single interpretation.

This article can be viewed as suggesting small but significant corrections to some views on interpretation that are commonly encountered in literary theory. The New Critics, Hirsch, and Fish all want to see meaning as a function of one argument. For the New Critics meaning depends on the text, for Hirsch on the author's intention. But neither of these computes. The text means nothing in the absence of rules to interpret it, and the author's intention is inaccessible until realized in some conventional way. By being explicit about the dependence of meaning on the rules of interpretation, or the conventions, one no longer has to argue about *which* rules or conventions determine *the* meaning of a text. The choice of a belief system to use is no longer an issue about "meaning" but an issue about the function of literature. Fish makes the opposite mistake. He discards the text and bases all on the reader or the interpretive community. Interpretations arise mysteriously, utterly unconstrained, out of interpreting activities. He supposes that interpretation can depend on only one thing, and recognizing its dependence on a system of beliefs, he is forced to banish what it is that is being interpreted. If we allow meaning to depend on two things, the text and a belief system, we are no longer forced into this implausible position.

Works Cited

Agar, Michael, and Jerry R. Hobbs. "Interpreting Discourse: Coherence and the Analysis of Ethnographic Interviews." *Discourse Processes* 5.1 (1982): 1-32.

Beardsley, Monroe C., and W. K. Wimsatt Jr.. "The Interpretation Fallacy." *The Verbal Icon: Studies in the Meaning of Poetry*, Ed. W. K.

Wimsatt Jr. Lexington, Kentucky: University of Kentucky Press, 1954.

Fish, Stanley. *Is There a Text in This Class?* Cambridge, Massachusetts: Harvard University Press, 1980.

Grice, H. P. "Logic and Conversation," *Syntax and Semantics*, Vol. 3, Peter Cole and Jerry Morgan, eds. New York: Academic Press, 1975.

Hobbs, Jerry R., and David A. Evans. "Conversation as Planned Behavior." *Cognitive Science* 4.4 (1980): 349-77.

Knapp, Steven, and Walter Benn Michaels. "Against Theory." *Critical Inquiry* 8.4 (1982): 723-42.

Knapp, Steven, and Walter Benn Michaels. "Against Theory 2: Hermeneutics and Deconstruction ." *Critical Inquiry* 14.1 (1987): 49-68.

Lakatos, Imre. "Falsification and the Methodology of Scientific Research Programmes." *Criticism and the Growth of Knowledge*, Imre Lakatos and Alan Musgrave, eds. Cambridge, England: Cambridge University Press, 1970.

Lewis, David. "Scorekeeping in a Language Game." *Journal of Philosophical Logic* 8 (1979): 339-59.

Mailloux, Steven. "Rhetorical Hermeneutics." *Critical Inquiry* 11.4 (1985): 620-41.

Stephen R. Schiffer. *Meaning.* Oxford: Oxford University Press, 1972.

Footnotes

1 – It is pleasant to speculate that this gratuitous attribution of human or humanlike agency is also the source of such phenomena as polytheism, hero

worship, and conspiracy theories.

2 – Since, presumably, it is better to be wrong than trivial, I take it that the generous reading of Knapp and Michaels is my original one.

3 – A set of people mutually believes a proposition if they all believe it, they all believe they all believe it, they all believe they all believe they all believe it, and so on, *ad infinitum* (Schiffer).

4 – I should mention that all of this is independent of consciousness. High-level goals, like "Sell this used car," tend to be ones we would be conscious of; very low-level goals, like "Use the word 'reliable' here," tend to be ones we would not be conscious of. AI in general has little to say about the experience of consciousness.

5 – In this assumption, we are taking positions on a number of controversial issues in AI and cognitive science, for example, the representability of knowledge in formal logic. These controversies, however, are not especially significant for the purposes of this paper. We could take other positions on these issues and construct a similar, though slightly different, framework and corresponding technical vocabulary to apply to the questions of interest in literary theory, and the results would be the same.

6 – See Lakatos. There is of course a significant problem concerning the epistemological status of "knowledge" acquired in this way, but because of their complexity, literary texts do not seem to be a good strategic locus for such an inquiry.

7 – I often use the terms "function" and "argument" in their mathematical sense. In the expression $quotient(60, 12) = 5$, $quotient$ is the function,

and 60 and 12 are its two arguments. Whether I intend these meanings or the ordinary sense of these two words should be clear from the context.

8 – He is also, of course, seriously underestimating the complexity of the *real* process of writing.