SRI International

A DESCRIPTIVE MODEL OF REFERENCE USING DEFAULTS

Technical Note 440

May 31, 1988

By: Douglas Appelt and Amichai Kronfeld

Artificial Intelligence Center
Computer and Information Sciences Division

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

The research reported in this paper was supported, in part, by the National Science Foundation Grant DCR-8407238. The views and conclusions expressed in this paper are those of the authors and should not be interpreted as a representative of the views of the National Science Foundation or the United States Government.

333 Ravenswood Ave. • Menlo Park, CA 94025
(415) 326-6200 • TWX: 910-373-2046 • Telex: 334-486

A Descriptive Model of Reference Using Defaults

Doug Appelt and Amichai Kronfeld

May 31, 1988

Contents

1 Introduction

In this article we try to answer the following question: how do we let our audience know what we are talking about? How, in other words, do a speaker and hearer form an agreement as to which entities are the subject of the conversation?

The question seems almost trivial: our hearer is expected to know what we are talking about because it is assumed that he understands the language we are using and, moreover, we have already *told* him what is being discussed. Nonetheless, the ease with which a native speaker can *refer* — i.e., indicate what entity is being discussed — is deceptive. Like other linguistic mechanisms, referring is easy to do but extremely difficult to explain. It is rather surprising that we hardly ever really tell our audience explicitly what we are talking about. Consider the following examples:

1.1 *Art* *Did you find Maya's shirt?*
Betsy *Yes. It was behind her toy chest.*

1.2 *Art* *You know, Jane's husband seems to be quite a romantic guy.*
Betsy *He is not her husband, you fool!*

In Example 1.1, the referring act is obviously successful, even though the girl named "Maya" is not the only person in the world so named; besides, she surely has more than one shirt. The referring act is successful, even though Art never *tells* Betsy explicitly what shirt he is talking about. In Example 1.2 the situation is even more complex: Betsy obviously knows exactly whom Art is talking about, and he knows that she does, despite the fact that what Art *told* her has to do with a different person entirely.

These two simple examples are neither unique nor exceptional. As a matter of fact, we hardly ever *tell* our hearers explicitly what we are talking about; we expect them to "figure it out." Indeed, in general we are very good at providing hearers with just enough information — no more and no less than necessary — to enable them to identify the subject of the conversation. How do we do that? And how can we teach a computer to do it?

Our goal is to provide an answer to this question. Such a response will ultimately take the form of a logical theory of reference that can serve as the basis of a computational model. However, implementation is not our main concern here, nor should the reader expect to find a blueprint for the constructing a specific system. Rather, our intention is to outline the general principles that ought to be incorporated in *any* specific implementation.

2 Internal Perspective and Standard Names

The referring problem is not new, of course, and each natural-language system has to face it in one guise or another. Reference is such an essential function of language that without it we could not communicate at all. But in almost all the systems we are aware of, the major mechanism for referring relies heavily on what we call *the standard-name assumption*, moreover, in most of them, the referring problem was approached from the *internal perspective*.

2.1 The Standard-Name Assumption

According to the standard-name assumption, *all* objects in the domain have standard names that are known to all participants in the discourse. In such systems, the act of referring is successful when (and only when) the machine associates the right standard name with the noun phrase. For

example, if a user types "The screwdriver is broken," referring to the object whose standard name is, say, the constant S_1, the referring act succeeds if and only if the machine associates that constant with the expression "the screwdriver." This one way of implementing the idea that the result of a speech act with successful reference is a singular proposition, with the standard name standing for the object itself.

This is, no doubt, quite useful if the system is expected to handle a small number of objects in a very limited way, but, as a general principle, the standard-name assumption is obviously too strong. It is unreasonable to assume that every object that can be talked about has a universally known standard name.

Moreover, within the framework of the logic of knowledge and belief, the standard name assumption has a rather undesirable consequence, namely, that agents can never be wrong about the identity of objects they can talk about. For example, under the standard name assumption, it would be impossible for Oedipus not to know that his wife is actually his mother. Since he can identify both, he must have a standard name for each, and the two standard names denote the same thing in all possible worlds, including, of course, all possible worlds that are compatible with what Oedipus knows. Hence, in all possible worlds compatible with what Oedipus knows, the mother and the wife are one and the same, which means, by definition, that Oedipus *knows* this identity. But surely an agent can be confused about the identity of an object while being perfectly capable of referring to it in a conversation.

2.2 Internal and External Perspectives

The act of referring is performed through the use and interpretation of noun phrases in a conversation. But we should be careful to distinguish between two perspectives — one *internal*, the other *external* with respect to the discourse. From the internal perspective, our main interest is the relation of coreference among symbols. From the external perspective, on the other hand, what interests us is the relation between symbols and the objects they represent. Consider the following exchange between Representative Louis Stokes and Assistant Attorney General Charles J. Cooper during the Iran-Contra hearings:

2.1 Stokes: *And lastly, when you and the Attorney General interviewed Colonel North, he was not under oath at that time, was he?*

Cooper: *No, he was not.*[1]

Although neither North nor the Attorney General — nor Cooper himself, for that matter — was under oath when they met, it is clear that, when Cooper says "No, he was not," he means North. How do we know that? How is the connection between the expressions "Colonel North" and "he" established? These are the typical questions that are asked from the internal perspective. Those that are asked from the external perspective, however, are different. How is the connection between the expression "Colonel North" and the *person* North established? What does it take for a hearer to recognize who the Attorney General is? When Stokes says "you," whom does he mean and how do we know what he means? This is what matters to us from the external perspective. Note that it is entirely possible that a native speaker of English would succeed in matching the expression "Colonel North" with the right instance of "he" without comprehending at all who Colonel North is. His success must be explained from the internal perspective, his failure from the external one.

Given the standard-name assumption, it is easy to ignore the external perspective entirely. Since standard names are simply labels, we tend to take the relation between the standard name

[1] *The New York Times*, June 26, 1987, p. 4.

and its bearer for granted. All that is left for us to do is to show how one symbol (the noun phrase) is associated with another (the standard name).

But the external perspective is indispensable, since, if we ignore it, we lose the basic rationale for the act of referring itself. Consider the following examples:

2.2 (a) *The farmer down the road owns a donkey named Buridan. He feeds it.*

(b) *The average farmer owns a donkey. He feeds it.*

(c) *If John owns a donkey, he feeds it.*

(d) *If a farmer owns a donkey, he feeds it.*

In Example 2.2(a), in contrast with Examples 2.2(b), the speaker has in mind a particular farmer and a particular donkey. But, from the internal perspective, this fact is of limited interest, since both "The farmer down the road," and "The average farmer" have equal potential for initiating anaphoric chains. Similarly, in Example 2.2(c), there is a particular owner that the hearer is expected to identify. No such identification is required for the interpretation of Example 2.2(d). Still, from the internal perspective all three — "John," "a farmer," and "a donkey" are treated equally: they are assigned *discourse entities* [19,39], which are basically "conceptual coathooks[2]" on which a hearer "hangs" subsequent noun phrases in the anaphoric chain. Whether the discourse entity corresponds to a real object or not is immaterial as far as the internal perspective is concerned.

But the question of whether the noun phrase corresponds to a particular object that the hearer is expected to identify is of prime importance for a natural-language system. Consider a speaker who is attempting to achieve something by means of language. Suppose, for example, that Luke Skywalker of Star Wars instructs his trusted robot C3PO to look for Han Solo's spaceship. It makes a great deal of difference to Luke whether it is understood that he has a particular spaceship in mind and, furthermore, whether the robot will be able to identify it. If the robot simply associates the phrase "Han Solo's spaceship" with the correct standard name and then switches itself off, Luke has not succeeded in his speech act. Any system that combines linguistic and nonlinguistic actions, and that is capable of cooperative behavior, must be able to talk about objects. It must distinguishes when a noun phrase has a referent in the real world from when it does not, when a particular type of knowledge of the referent is required from when it is not, when knowledge of the referent is presupposed from when it should be actively sought. Without the external perspective, we do not even have a basis for asking these questions.

Our main objective, then, is to develop a referring model from which the standard-name assumption can be eliminated and in which the external perspective receives the attention it deserves.

3 The Individuation Principle

From the standpoint of the external perspective, we have the object the speaker is thinking of, we have the referring expression used by him, and we ask how a hearer makes the connection. Once the question is formulated in this manner, however, it becomes readily apparent that it masks a more general problem. Never mind how the hearer recognizes the connection between a referring expression and an object. How is this connection established in the first place? What does it mean to say that the speaker has a particular object in mind? Does it mean that he is able to identify that object when he sees it? Does it mean that he knows something that is true of that particular

[2]The term "conceptual coathook" is attributed to William Woods.

object and no other? Such questions lead to what may be called the problem of *reference*, which can roughly be phrased as follows: "How can thoughts (and sentences that articulate them) be *about* objects?" The problem seems simple enough, but, as was the case with the *referring* problem (i.e., how do we let our audience know what we are talking about?), deceptively so. In effect, a solution to the problem of reference would illuminate the general mechanism that enables the mind (and, derivatively, language) to represent the world for us. It is not surprising, therefore, that the problem of reference has occupied a central position in the philosophical debate that has been going on since the very beginning of this century.

Our chosen philosophical approach to reference is what we call the *descriptive* research program, which can be characterized as founded on two central ideas. The first idea is that to refer to an object — in thought or in speech — is essentially to have or invoke a mental representation of that object. The second idea is that the relation between a thought and the object it is about is that of denotation, which in turn is a function of descriptive content. Thus, the crux of the descriptive program is that reference is entirely a matter of associating a mental state with descriptive content.

The descriptive approach hardly enjoys unanimous acceptance. As a matter of fact, in the past two decades or so it seems to be on the defensive as a new approach (the *causal* approach) has emerged. A full discussion as to why the descriptive framework is superior would take us too far afield (see [26,29,27]). Here we shall argue only for what is called the *individuation principle*, which we consider to be a strong motivation for adopting the descriptive approach. Roughly, this principle states that, (1) the ability of an agent to think of an object depends on the agent's having a *presentation mode* of that object, and (2) for each possible world compatible with the agent's beliefs, that presentation mode determines one and only one object. In other words, we want to argue that the individuation of objects in the agent's mental life is done through presentation modes. This conclusion, as we shall see, shapes the basic principles of our model.

We shall start by stating a premise that we take to be trivially true. we call it the *trivial principle*:

Trivial Principle: It is impossible both to hold and not to hold the same belief.

It is important that the triviality of this claim be appreciated. This is not a characterization of our rationality. We do not actually contend that it is impossible to hold both a belief and its negation. Although it is not particularly recommended, one can certainly hold the belief that P and, at the same time, hold the belief that not P. A belief and its negation are two distinct beliefs and, if one chooses to hold both, one is free to do so. But an agent cannot both hold and not hold the *same* belief any more than a beer bottle can be both in the cooler and not in it.

Now, what is the content of a belief about a particular object? Let us take a concrete example. Suppose Ralph believes of Wiley that he is a spy. The content of this belief, according to the causal theory of reference (at least in its original form), is the singular proposition:

3.1 $\mathbf{SPY}(Wiley)$.

But, because of the *trivial principle,* the singular proposition cannot be the *complete* specification of the content of Ralph's belief. Suppose that, on a certain occasion (say, on the beach), Ralph points to Wiley and says "I believe this man is a spy." Suppose that, at another time (say, in a supermarket), Ralph points to Wiley and says "I do not believe this man is a spy." Let us assume that Ralph is sincere on both occasions and that his only problem is his failure to recognize Wiley in the supermarket as the man he saw earlier on the beach. If the *complete* content of Ralph's belief is a singular proposition, Ralph's first utterance shows that he holds a belief whose content is

expressed by (3.1), while his second utterance shows that he does *not* hold that same belief. But, according to the *trivial principle,* this is impossible.

If the singular proposition is not the complete content of Ralph's belief, some element of content is missing. Let Ralph's mode of presentation (of Wiley) be *by definition* that missing element. We have not said at this point what a mode of presentation is. What we do know is that a mode of presentation, in the sense described above, is necessary regardless of what one's theory of the content of *de re* beliefs is. The *trivial principle* simply requires it.

Now, whatever we consider modes of presentation to be, they must satisfy the following condition, which we call the *basic constraint*.

> **Basic Constraint:** For every mode of presentation M_1 and M_2, if $M_1 = M_2$, then, if Ralph believes Wiley to be a spy under M_1, he also believes Wiley to be a spy under M_2.

We take the *basic constraint* to be as self-evident as the *trivial principle*. It is nothing more than an instantiation of Leibnitz's law: if two things are identical, whatever is true of one is true of the other.

Now, from the *basic constraint* on any theory of presentation modes it is possible to derive another principle, which we shall call the *individuation principle:*

> **Individuation Principle:** If M is a mode of presentation under which Ralph believes Wiley to be a spy, then, in each possible world that is compatible with Ralph's beliefs, one and only one object is presented to Ralph under M.

What the *individuation principle* means is that modes of presentation — whatever they are — must carry out an individuation function within one's network of beliefs. In other words, if M, the mode of presentation under which Ralph believes Wiley to be a spy, is a concept, it is an individual concept — instantiated by a unique object in each possible world compatible with Ralph's beliefs. If M is a description, it is a definite description — denoting a unique object in each world compatible with Ralph's beliefs. If M is a causal chain, it "determines" a unique object in each world compatible with Ralph's belief, and so on with regard to anything a theory of presentation modes might suggest.

The argument for the *individuation principle* is, in essence, identical to the original motivation for modes of presentation. Let us suppose that the *individuation principle* is false and, to make the argument more concrete, let us assume that, in our theory of content, modes of presentation are nonindividuating *concepts*. Suppose, for example, that the mode of presentation under which Ralph believes Wiley to be a spy is *man-on-the-beach*. Since Ralph considers Wiley to be a man on the beach, this concept is instantiated by at least one individual in every world compatible with Ralph's beliefs. As this is a nonindividuating concept, however, it is certainly possible for Ralph to believe at another time, that he sees a different man on the beach; Ralph has no opinion as to whether or not he is a spy. Thus, on one occasion Ralph thinks to himself "I believe this man on the beach is a spy," whereas at another time he thinks "I do not believe this man on the beach is a spy." But suppose that, on both occasions, the man really is Wiley (although Ralph does not realize that). If the complete content of Ralph's belief is

3.2 SPY(*Wiley*) [under mode of presentation: *'man-on-the-beach'*],

we again find Ralph both holding and not holding the same belief. But this would be impossible, since it too violates the *trivial principle*.

We have borrowed this argument from Schiffer [35], who uses it to show that nonindividuating concepts cannot be modes of presentation. But the same argument can easily be generalized to show that the *individuation principle* must be right no matter how one interprets modes of presentation. The schema of the argument is as follows. Let us assume that the *individuation principle* is false. There are then possible worlds compatible with Ralph's beliefs in which two distinct individuals are presented to Ralph under M. Nothing prevents Ralph from believing that one of them is a spy, while simultaneously *not* believing that the other is. But, since Ralph may fail to recognize that the man is Wiley in both cases, we have a situation in which Ralph believes Wiley to be a spy under M, while *not* believing Wiley to be a spy under M. This contradicts the *basic constraint* as well as, of course, the *trivial principle.* Hence, our hypothesis that the *individuation principle* is false must be incorrect.

Our contention, therefore, is that the *individuation principle* renders the descriptive research program more promising. Modes of presentation are needed no matter what one's theory of the content of beliefs is, and modes of presentations, whatever they may be, must individuate objects for agents. Given these facts, the ease with which the notion of descriptive content can accommodate them, and the difficulties of incorporating a theory of modes of presentation within the causal approach, it seems to us that the descriptive approach is the logical choice.[3]

4 Referring as Planning

The act of referring is normally done by means of noun phrases in conversation. However, not all noun phrases are intended to be used in this manner, not even those that have the form of a definite description. For example, in the sentence "The whale is a mammal" (uttered, say, in a biology class), the speaker is making a general statement in which no referring relation is presupposed between the noun phrase "the whale" and any individual whale. Let us reserve the term *referring expressions* for those instances of noun phrase usage that are intended to indicate that a *particular* object is being discussed. Note that one and the same noun phrase may sometimes function as a referring expression, other times not. While going whale watching, for example, a child may point to a whale and comment on what a big fish it is. His mother can correct him by saying "The whale is a mammal," meaning that that particular whale is a mammal. Here, the noun phrase "the whale" is clearly used as a referring expression, in contrast to the above example.

Thus, whether or not a particular noun phrase is a referring expression depends on the way it is intended to be interpreted. A theory of referring, therefore, is not a theory of language but rather of language *use.* In general, theories of language use (that is to say, *pragmatic* theories) specify and explain the ability of humans to employ language for some purpose. Consequently, an account of referring should specify and explain human competence in using referring expressions to achieve particular goals. Now, pragmatic theories have concentrated on two complementary aspects of language use. The first, which is at the heart of Grice's theory of meaning, is this: when we use language, we typically achieve some of our goals by making our audience recognize our intentions to achieve them. For example, I can succeed in congratulating you simply by making you recognize my intention to do so. Once you have recognized my intention, you are thereby congratulated and nothing else is necessary. This is a unique feature of communication, as Grice [16] was the first to notice. The second aspect of language use, which is a central element in Searle's speech act theory is this: we make our audience recognize our intentions by following mutually known rules that determine what the utterance of a particular expression *counts as.* For example, underlying

[3]The description program, no doubt, has its own difficulties. Kronfeld discusses ways to overcome some of these difficulties [26,27].

the recognition of an intention to pay one's debt is a rule that is mutually known by both speaker and hearer; this rule specifies that the utterance of "I hereby promise to pay my debt" *counts as* placing the speaker under the obligation of paying his debt [36].[4]

These two general principles of language use determine the structure of pragmatic theories. For any communication act, such theories should state precisely what relevant speaker's goals are involved, and on what basis a speaker expects and intends these goals to be recognized. Moreover, since the relation between language use and a speaker's goals is what needs to be explained, it is natural, within the context of computational linguistics, to consider language use as a *planning* problem [1,4,8,11]. What underlies the *generation* of an utterance is a plan (constructed by the speaker) to achieve certain goals through available means (linguistic or otherwise). The *understanding* of the utterance involves the hearer's recognition of the goal, as well as of the plan itself (or perhaps just a part of it). By regarding language use as a special case of planning, we are provided with a large array of computational tools that have been developed within AI in recent years. Moreover, since planning is a special form of rational behavior, the justification of rules for language use can be grounded upon a general theory of rationality [10,20,21].

Such a computational approach to language use governs the referring model we are after. A plan-based account of referring is an integral part of a plan-based theory of speech acts. At a certain point in the planning of a speech act, it may become obvious that, as a precondition for further steps in the plan, the speaker must make the hearer identify a particular object as being relevant to the conversation. To achieve this goal, an act of referring then becomes necessary. Thus, our computational model of referring must show how the successful use of a referring expression in a given context is due to the solving of a planning problem — given also a goal, various rationality assumptions, and relevant linguistic institutions.

In sum, what we have been saying so far is this: a pragmatic theory of referring is one that specifies and explains human competence in using *referring expressions* to achieve certain goals. Since the relation between referring expressions and a speaker's goals is what must be explained, it is natural to consider referring as planned action. This in turn requires showing how the use of referring expressions is systematically related to changes in both the hearer's and speaker's mental states.

In order to do that, however, we need to specify the goals that typically motivate an act of referring. These are the goals that characterize the changes in the hearer's state of mind that the speaker is trying to accomplish.

5 The Literal Goal and Discourse Purpose of Referring

In performing one and the same speech act, a speaker may have many distinct goals. For example, by uttering "The house is on fire!" a speaker may intend to inform the hearer that the house is on fire, scare the hearer half to death, and/or make the hearer leave. Not all such goals (and the intentions to satisfy them) are relevant to us. What we are seeking are *communication* intentions and goals — goals that are intended to be *recognized*, or, more precisely, to be achieved at least in part, through their recognition.

This view of communicative intentions originates with Grice's analysis of the concept of meaning [16]. But much research in computational linguists, though obviously influenced by Grice, has nevertheless stressed the role of intention and goal recognition in discourse, quite independently of a theory of meaning. Allen's dissertation [1] and his subsequent work with Perrault and Cohen

[4] This is the rule that defines the institution of promising [Ibid., p. 60]. Searle calls such rules *constitutive rules* [Ibid., pp. 33–42.]

[3,2,33], for example, emphasize the importance of goal recognition for inferring the speaker's plans. So does Sidner in her own work [38,37], as well as in her collaboration with Grosz [18]. These authors show how the recognition of what the speaker intends contributes to discourse coherence and comprehensibility, and is essential for the hearer's generation of an appropriate response. However, not all goals that are intended to be recognized are alike.

Sometimes, the recognition of a goal is enough for its satisfaction. For example, if my objective is to congratulate you, we can succeed if (and in this particular case, only if) you recognize my intention to do so. Once you have recognized my intention, you are thereby congratulated. We call such a goal, whereby recognition is sufficient for success, the *literal goal* of the speech act. In addition, there are what Grosz and Sidner call *discourse purposes*, which are the goals that underlie both the choice to engage in discourse in the first place (rather than in a nonlinguistic activity), and the choice of a particular propositional content to be expressed [18]. In the case of congratulating, literal goal and discourse purpose are one and the same, but this is the exception rather than the rule. For example, suppose Art asks Ben to take out the garbage. The literal goal of the request is roughly to make Ben *realize* what he is asked to do. The discourse purpose is to make him actually *do* it. Note that, unlike the case of literal goals, the recognition of the discourse purpose (though important for the success of the speech act) is in general not enough for the purpose to be achieved. Ben may very well recognize that Art's purpose is to make him take out the garbage, but, alas, this in itself is no guarantee of his cooperation. So the recognition of literal goals is sufficient for their success (sometimes it is also necessary), while the recognition of discourse purposes, as distinct from literal goals, is neither sufficient (it does not guarantee cooperation) nor necessary (Ben may take the garbage out without realizing that Art wants him to do so). Rather, given certain assumptions about the disposition of discourse participants, it is *rational* to expect that the hearer's recognition of discourse purposes will enhance the speaker's chances of achieving his goal.

Literal goal and discourse purpose are obviously not independent of each other: Art's expectation that Ben will actually carry out the garbage dependes partially on Art's expectation that Ben will understand his request. This is, indeed, the basis for our utilization of language as a means for achieving some of our objectives; it is also the foundation for a plan-based theory of speech acts. Since we accept the plan-based approach and regard referring as a speech act, we must specify the literal goals and discourse purposes that typically motivate the use of referring expressions. Needless to say, the literal goal and discourse purpose must be characterized in a way that is useful for the referring model.

5.1 The Literal Goal of Referring

Let us begin with Grice's theory of communication intentions. The backbone of Grice's analysis of meaning *something*, and the foundation for his entire approach, consists of three intentions that are supposed to be both necessary and sufficient for a speaker to mean anything [16]:

Intention 1: S intends to produce an effect in hearer H.

Intention 2: S intends H to recognize his Intention 1.

Intention 3: S intends that Intention 1 be satisfied by means of H's recognition of Intention 2.

In other words, if S is to mean anything at all, he must intend to produce an effect in H *by means of H's recognition of this intention.*

As a theory of meaning, Grice's account got into trouble of various sorts, and a series of counterexamples indeed forced him to modify the three intentions quite extensively [15]. However, our interest here is not in a theory of meaning per se. Whether Grice's account is, or could be

made to be, a correct analysis of the concept of meaning is debatable. Apart from that question, however, his original insight about the nature of communicative intentions and goals is still valid. So if Grice is right, referring intentions must include communicative intentions, whose structure corresponds rather closely to the Gricean picture. What we need to find, therefore, is the kind of referring intention that would require no more than to be recognized for its satisfaction. This intention would provide the *literal goal* of the referring act.

Once we adopt the descriptive approach to reference, such an intention is not hard to find. Consider what an intuitive account of referring looks like from a descriptive point of view. A speaker has a mental representation denoting an object. By using a noun phrase that is intended to be interpreted as a linguistic representation of that object, the speaker intends to invoke in the hearer a mental representation denoting this very same object. But note that such an intention has precisely the Gricean quality we are looking for. Once the hearer recognizes the intention that he have a mental representation denoting the same object that the speaker has in mind, the hearer *does* have such a representation. For example, if I recognize your intention to impart to me a representation of whatever object you are talking about, then I *do* have a representation of that object under the presentation mode: *the object the speaker is talking about*. In other words, *the mere recognition of the intention to represent an object is enough to produce a presentation mode of that object*. A central referring intention, therefore, is the intention to invoke a representation of a particular object in the hearer by means of his recognition of this intention. The goal of satisfying this intention is the literal goal of referring.

We can also describe the literal goal of referring differently in terms of features of noun phrases. As Grice points out

> Characteristically, an utterer intends an audience to recognize ...some "crucial" feature F [of the utterance], and to think of F ...as correlated in a certain way with some response which the utterer intends the audience to produce. [14, p. 163]

If the property of being a referring expression is taken to be a feature of some noun phrases, and if this feature is correlated with the invocation of a mental representation in the hearer's mind of whatever object the speaker is talking about, we can then say that the literal goal of referring is to have the hearer recognize that an utterance of a noun phrase is to be interpreted as a referring expression. Of course, noun phrases that are used as referring expressions have other features relevant to the act of referring. For example, when used as a referring expression, the noun phrase "the gray whale" has a feature that is conventionally correlated with invoking in the hearer a representation of a gray whale. Thus, the literal goal of referring is not merely to invoke a representation of an object under the presentation mode *the particular object the speaker wishes to talk about*, but to invoke a representation of that object with properties that correspond to certain features of the noun phrase. In other words, if the speaker's intention is to refer to an object by using a noun phrase, the literal goal of his referring act is to make the hearer generate a presentation mode denoting that object by virtue of the hearer's recognition that the noun phrase is to be interpreted as a referring expression. Computationally, this means that when it has been recognized that the noun phrase is to be interpreted as a referring expression, the hearer generates a single presentation mode, namely,

5.1 *The one and only x, such that x is the ...the speaker wants to say something about.*

The ellipsis in (5.1) is to be filled by the descriptive content of the referring expression. For example, if the noun phrase is "the gray whale," the corresponding element in the hearer's newly created presentation mode is

5.2 *The one and only x, such that x is the gray whale the speaker wants to say something about.*

Naturally, some noun phrases do not have any descriptive content conventionally associated with them (e.g., "this"). In such cases, the generated presentation mode is simply the *object* the speaker wants to say something about. Thus, it is important to note that the descriptive content of referring expressions is not at all necessary for initial individuation of the referent. The literal goal of referring can be achieved even if the descriptive content denotes nothing, is applied incorrectly, or is lacking altogether.

For complex reasons, some of which will be discussed later on, the most important abstract data structure in our model is not a presentation mode, but a cluster of such modes. We call such clusters *individuating sets*. Roughly, an individuating set is an equivalent class of mental representations: if we take the class of all representations that are believed by an agent to denote one and only one thing (i.e., representations of the form "the one and only thing x, such that $F(x)$"), and if we cluster together all representations that are taken as denoting the very same object, each such cluster is an individuating set.

Given the concept of an individuating set as an abstract data structure representing a particular object for an agent, the computational interpretation of the literal goal of referring is that the hearer generate an individuating set containing the single presentation mode, as discussed above. The need for this should become obvious in the next section.

5.2 The Discourse Purpose of Referring

Success in achieving the literal goal of referring is hardly enough. The hearer must still *identify* the referent. Referent identification is the point of referring, but the concept of identification itself is so vague and ambiguous that it is of little help for a computational approach. Let us begin, therefore, with two important distinctions.

First, the speaker's sense of "identify" is quite different from that of the hearer. A speaker is said to identify an object *for* or *to* a hearer, while the hearer is said to identify that object if the referring act is successful. In what follows we shall concentrate on the hearer's sense, regarding the speaker's sense simply as the plan to achieve the hearer's identification of the referent.

Second, identification, as the discourse purpose of referring, should be carefully distinguished from identification in the sense of *knowing who* someone is. The former may be said to be a *pragmatic* notion of identification. The latter is an *epistemic* one. Although the pragmatic notion of identification is also connected with knowledge (in the sense of knowing whom the speaker is talking about), the two are quite distinct. To illustrate the difference, consider the following case. Two policemen on the beat, Art and Ben, discover clear (and fresh) signs of a break-in. "Quick!" says Art. "He must be close by!" What Art means, of course, is "Quick — the burglar must be close by!" Neither Art nor Ben knows who the burglar is, moreover, they cannot "identify" him if identification is interpreted epistemically. From a pragmatic point of view, however, there is a clear dichotomy: if Ben makes the connection between "he" and "the burglar," he has identified whom the speaker is talking about. Otherwise he has not.

Not having to deal with the complex concept of *knowing who* is certainly a relief, but we are hardly in the clear as yet. The pragmatic notion of identification seems as elusive as its epistemic counterpart. Consider the following examples:

5.3 (a) *Look at* this fellow *move! He must be doing at least a five-minute mile!*

(b) *Do you remember* the little playwright *who used to hang out with us in the sixties? He has just won the Pulitzer prize.*

(c) *Tell me what other plays were written by* Shakespeare.

5.4 (a) Your friend *has just won $10,000.*

(b) My friend *has just won $10,000.*

(c) A friend of mine *has just won $10,000.*

When we try to formulate the conditions under which a hearer can be said to identify correctly the referents in these examples, we find that in each case there is a different standard. In Example 5.3(a), the hearer is asked to identify the runner, in the sense of locating him in his visual field, but in Example 5.3(c) visual identification of Shakespeare is clearly not required, although the hearer is still expected to identify the referent in some other way. In Example 5.3(b), locating the playwright in one's visual field is also not called for, but the requirements for correct identification are surely different from Shakespeare's case. In the latter case the hearer needs to associate the name with a bulk of shared cultural knowledge. In the playwright case, a connection between the description and perhaps a memory image is required.

Not only do methods of identification differ from utterance to utterance, but there are also differences in expected *degrees* of identification (analogous, perhaps, to variations in degrees of illocutionary force). In Example 5.4(a), the speaker would usually expect the hearer to know (or inquire) which friend the speaker is talking about. In uttering Example 5.4(c), on the other hand, the speaker probably intends that the hearer does not have to know which friend the speaker means. All that the hearer is expected to do is to note the existence of a particular individual, a friend of the speaker, who is lucky enough to have won such a large sum of money.[5] The identification requirements in 5.4(b) may be similar either to 5.4(a) or 5.4(c), depending on contextual factors and the purpose of the conversation.

Obviously, there is not just one "correct" method of referent identification. Still, we can look for a way to express the differences as variations on a single mechanism. If the reason for referent identification is the need to establish mutual agreement as to which object is being talked about, then a necessary condition for successful referring is that the hearer understand the ground rules for reaching such an agreement. In general, such ground rules follow from the propositional content of the illocutionary act itself, from general knowledge about the discourse (in particular, speaker's goals), and from principles of rational behavior. For example, according to any theory of action it would have to follow that, if the speaker asks the hearer to pick up an object, the hearer can comply with the request only if he knows the position and orientation of the object he is expected to manipulate. This fact implies a rather specific mode of identification: locating the object in one's immediate vicinity. Different ways to identify the referent are similarly derived in other circumstances. What is important for our model at this stage is this: if we accept the descriptive approach, we are committed to expressing or formalizing the rules for pragmatic identification *in terms of presentation modes*. In other words, the assumption is that a necessary and sufficient condition for a hearer to identify whom or what the speaker is talking about is that the hearer come to possess one or more appropriate presentation modes of the referent. Never mind, for now, what an "appropriate" presentation mode is. The important point is that pragmatic identification

[5] We take the phrase "a friend of mind" in 5.4(c) to be a referring expression because there is a clear indication that a particular individual is being talked about

is taken to be *entirely* a matter of having a representation whose descriptive content denotes the referent and that is appropriate in a sense as yet to be explained.[6]

How can we begin to represent such a view in our model? If the hearer is cooperative, his pragmatic identification of the referent begins once the literal goal has been achieved. At this point, the hearer can be said to possess a single presentation mode that is guaranteed to denote whatever the speaker has in mind. What happens next depends on circumstances, but it would be wrong to assume that there is always a unique presentation mode by which identification succeeds or fails. For example, suppose that, during a discussion of German theater, I mention Bertold Brecht to you. I believe that you have an individuating set of presentation modes associated with the name "Bertold Brecht" that allows you to understand whom I am talking about. Nevertheless, I hope that a significant segment of your individuating set would be fairly similar to the one I have, although there is usually no *particular* presentation mode that must be there, nor, of course, do I expect your individuating set to be identical to mine. What I expect, rather, is that your individuating set satisfy certain *constraints*: it should include, for example, a sufficient number of individuating facts concerning important plays by Brecht, but the list does not have to correspond exactly to my list (nor does it have to be exhaustive).

Thus, instead of representing pragmatic identification as a relation between a hearer and a presentation mode, it is better to represent it as a relation between the hearer and a pair of formal entities: an *individuating set* and a set of *identification constraints*. If the literal goal of referring is to make the hearer generate an individuating set that contains a presentation mode of the referent, then the discourse purposes of referring is (1) to make the hearer understand what identification constraints are operative, and (2) to have him apply these constraints to the newly generated individuating set. Possible identification constraints may include the following requirements:

- That the relevant individuating set contain a *perceptual* presentation mode, to be acquired now or in a later time.
- That the hearer be able to merge the newly generated individuating set with one that he already had prior to the conversation. (Further identification constraints may be necessary to ensure that the latter is adequate for current purposes, as discussed below).
- That the hearer be able to connect the new individuating set with one generated earlier in the conversation. In particular, this includes cases such as "An old girlfriend of mine got married yesterday. *The lucky man* met her only three weeks ago."[7]
- That the relevant individuating set contain one or more presentation modes that are *privileged* relative to the goals of the speaker. Suppose Art asks Ben: "Do you think *Ronald Reagan* will send the marines to invade Nicaragua?" Ben may have a very rich individuating set for Reagan. He may know, say, each and every detail of Reagan's Hollywood career. Yet, if this individuating set does not contain a presentation mode such as *current president of the United States* or *commander in chief*, pragmatic identification has not been accomplished.
- The null identification constraint, in which success of the literal goal is already sufficient for pragmatic identification. "*An old friend of mine* once told me that ..." is an example. Since

[6] Of course, this does not mean that the appropriate presentation mode required for pragmatic identification can always be verbalized as a definite description. When I want you to locate an object in your vicinity, I want you to see it (or touch it, etc.), and perception is as good a source of presentation modes as any. But obviously, I do not necessarily expect you to translate what you see into words.

[7] We intentionally refrain from using the term "anaphora" here. Although anaphora resolution is essential for pragmatic identification, it belongs to the *internal* perspective, and is a much more general phenomenon. The two should be kept apart, at least conceptually.

the speaker merely indicates the existence of an old friend whose identity is irrelevant, once the literal goal has been achieved no further identification is needed.

This view of pragmatic identification in terms of individuating sets and constraints upon them is merely a tentative suggestion at this stage. From a computational standpoint, this approach awaits careful specification and formalization of identification constraints, as well as an explanation as to how they are derived and satisfied by a hearer — not an easy task by any means. But an important advantage of this view can already be pointed out: it enables us to eliminate the *standard-name assumption*: an identification constraint may be satisfied even though the speaker and the hearer does not share a standard name that denotes the referent.

In summary, there are two goals that motivate an act of referring: (1) the literal goal is to activate in the hearer's mind a presentation mode denoting the referent, by means of the hearer's recognition of this goal. (2) the discourse purpose of referring is pragmatic identification. This second goal is achieved, if the hearer is cooperative, in two steps: first, the hearer derives the appropriate identification constraints, and second, he attempts to identify the referent by applying those constraints to the set containing the initial presentation mode.

6 A Logical Model of Referring

In this section we explain how to construct a model of referring based on (1) the descriptive model outlined in the previous section, and (2) a general theory of speech acts. It is important that any theory of referring be integrated with a theory of speech acts because referring is almost always an essential component of the latter. We view speech act understanding as a process of attitude revision in response to an utterance. Since referring is a type of speech act, similar principles of attutide revision should account for the understanding of referring expressions as well.

We believe that nonmonotonic reasoning is essential to a perspicuous and computationally adequate description of reference. Perrault [31] argues that there are no interesting or important effects of an utterance that hold in every instance of its performance. However, because it is obviously true that speech acts do have an intuitively clear set of conventional effects, he proposes to represent the conventional effects as defaults that can be overridden by contextual elements relevant to the performance of any particular speech act. The situation is much the same with regard to reference. One can inform [4] or request [9] through a referring expression, and even refer ironically, whereby the description used to refer is mutually believed to *not* hold of the referent. However, in the typical case in which he uses a referring expression, the agent believes that the description he uses is actually a property of the intended referent, and that knowledge of this property is sufficient in the current context for the hearer to know which object the speaker is talking about. We believe that the most perspicuous model of referring can be constructed in a manner analogous to a general theory of speech acts: conventional elements of the act are stated as default rules that can be defeated in a particular context.

It is also important to integrate a theory of speech acts with the theory of reference because it is easy to find examples in which both must be integrated so as to account for observed behavior. Consider, for example, a speaker who utters the following sentence during a cooperative task of assembling a toy water pump [13]:

You now have one red piece remaining. (1)

Suppose the hearer of this sentence believes that he has exactly one piece remaining and it is pink. At this point, the hearer seems to have a choice. He can either accept the speaker's description

"red piece" literally, and believe the speaker is either insincere or mistaken about the predication that it is the only piece remaining. On the other hand, the hearer could believe that the speaker is talking about the remaining pink piece, and making a predication that is consistent with the hearer's beliefs, but the description employed by the speaker to refer to it was the result of a mistaken belief. There is, of course, no general answer to the question of how a hearer will revise his beliefs in response to a speech act on any given occasion. However, whatever theory of reference is proposed, it must not focus exclusively on the referring expression, but rather on the speech act as a whole. Whatever theory is proposed should be able to accomodate both outcomes of the belief revision process.

Appelt and Konolige [5] have proposed a theory of speech acts based on a nonmonotonic theory called Hierarchical Autoepistemic Logic (HAEL). This general framework is well suited to the task of formalizing a theory of referring for two reasons. First, rather than postulating that an agent's belief consists of a single monolithic theory, beliefs are rather composed of a number of separate theories that capture different aspects of the agent's knowledge about a particular situation. Each of these theories may incorporate default rules. Furthermore, these theories are arranged in a partial order that determines how information contributed from each of them is combined. Also, each theory inherits the information from the theories below it in the hierarchy. The fundamental advantage of representing knowledge in this hierarchy of theories is that it provides a means of staing prioirities among default rules. Defaults associated with theories lower in the hierarchy take precedence over those associated with theories at higher levels. As we shall see, this ability to state preferences among default conclusions plays an important role in the formalization of a theory of reference.

The second reason for the choice of HAEL as a suitable formalism is that it is has a number of properties that make it especially suitable for use in a computational system. First-order nonmonotonic logics are inherently undecidable in general, because theoremhood for first-order logics is only partially decidable. HAEL is no exception in this regard. However, HAEL does have some properties that make it particularly amenable to computational implementations — properties that are not shared with other nonmonotonic theories.

Typically, nonmonotonic logics have the problem of multiple extensions (default logic, [34]) or stable expansions (autoepistemic logic, [6]). This means that a wff is a theorem of the logical theory only with respect to a particular extension, and theoremhood with respect to the theory as a whole is not well-defined. The proliferation of extensions is a consequence of the ordering of the application of the defaults with mutually inconsistent consequences, but there is no way within these logics to express *which* ordering is actually intended. Perrault encounters some serious difficulties [5] attempting to formalize a theory of speech acts within such a logic [31], with rules constrained so that the theory has only a single extension. Any consistent HAEL theory, however, has only one extension, membership in which is decidable if all of its constituent theories are decidable. A more complete formal description of HAEL, along with a discussion of its important computational properties, can be found in Appelt and Konolige [5] and Konolige [23].

6.1 An Overview of the Referring Model

Before delving into the formal details of the referring model, we shall outline here the model's general structure, as well as some of the general characteristics that we believe any model of referring must posess. We feel that a referring model should characterize not only the manner in which beliefs and intentions are affected by a speech act in the most common situations, but also that, when some of the conditions that are usually assumed to hold (e.g. the speaker and hearer do not mutually believe that the referring description holds of a unique individual) are violated, the model should

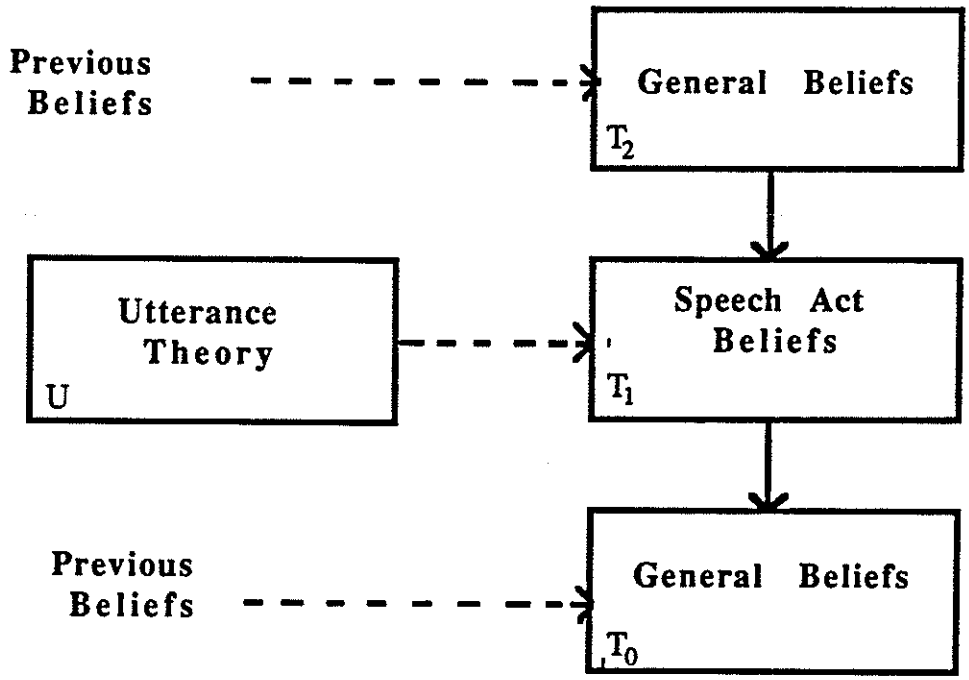

Figure 1: A HAEL Theory of Atomic Propositional Speech Acts

then too make some predictions about what the result of the speech act will be, or at least what it *can* be, given various assumptions about the speaker's and hearer's belief revision processes.

The general approach will be to divide the speaker's and hearer's knowledge into a number of *logical theories,* i.e. sets of sentences closed under rules of inference. The theories will be ordered in a hierarchy according to a partial ordering "$\prec$." Intuitively, $T_i \prec T_j$ means that theory T_i takes precedence over theory T_j, in that any default conclusion derived in T_j can be overridden by evidence to the contrary in T_i, but not vice versa. Figure 6.1 outlines how an agent's knowledge about speech acts relates to his general beliefs. The information present in the utterance is contained in the utterance theory u. Theory T_1 represents the agent's knowledge about speech acts and communication. Beliefs are transfered from theory u to T_1, provided that they are consistent with the beliefs in T_0 that persist from the situation that existed prior to the utterance. Beliefs derivable in theory T_1 take precedence over those derivable in T_2, so that this model provides a means of stating which beliefs can be defeated by an agent's speech act and which cannot, by specifying to which theory, T_0 or T_2, beliefs are transfered from the previous state. The solid arrows in Figure 6.1 represent the partial-ordering relation. The theory of speech acts explains how information about speech acts relates to the general beliefs of the speaker and hearer. Exactly how these beliefs are updated — i.e., the details of which beliefs are transfered from T_0 and T_2, is an interesting empirical question, but one that we are not prepared to address at this time.

If the model of Figure 6.1 is to be extended to model speech acts with referring expressions, instead of treating the propositional content of the speech act as an atomic proposition, it needs to be elaborated along the lines indicated in Figure 6.1. In Figure 6.1, the basic model is augmented with theories T_2 (attentional state) and T_3 (description adoption). The role of these additional theories is to model the integration of the beliefs resulting from the referring expression with those

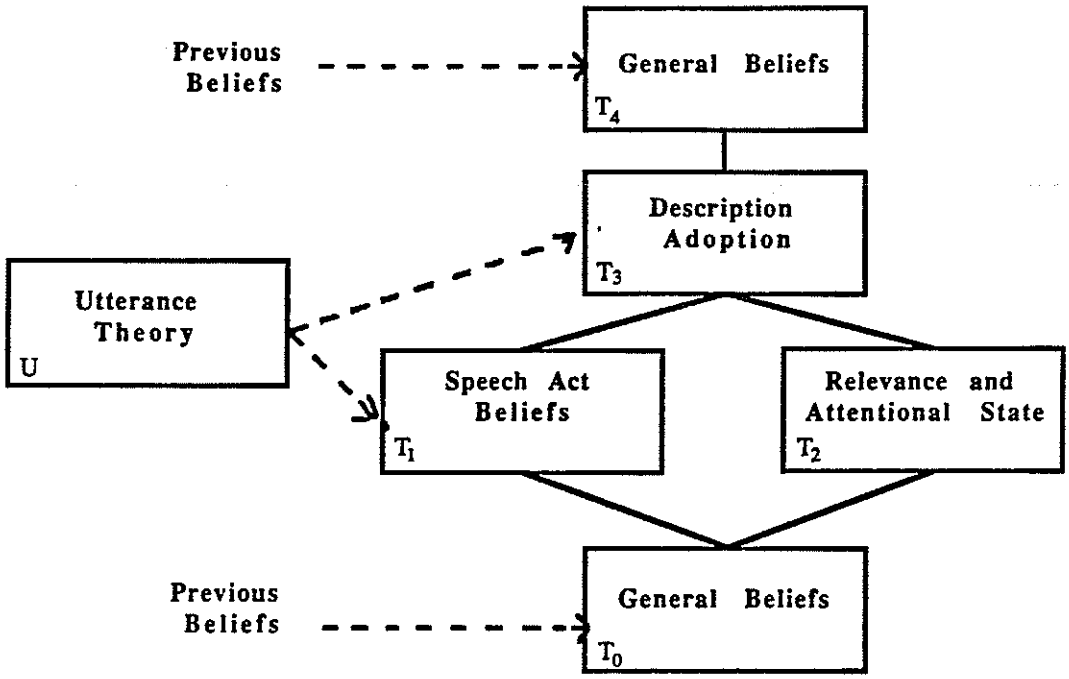

Figure 2: A HAEL Theory of Speech Acts with Referring Expressions

derived from the propositional content of the utterance and the hearer's general knowledge.

The utterance theory u will contain not only the utterance's propositional content, but also information about the kinds of descriptions used in the utterance's referring expressions. The attentional state theory, T_2, describes the speaker's and hearer's mutual belief about what individuals are relevant to the discourse, whilethe description adoption theory T_3 describes how various aspects of the description used by the speaker are adopted by the hearer, so that they are consistent with theories T_1 and T_0. In the next two sections, we explain what the contents of these theories are and how the nonmonotonic rules in them account for the way a hearer's beliefs change when the actual situation of the utterance differs from the "typical" situation.

6.2 A Theory of Speech Acts Based on HAEL

We shall now begin to flesh out the model of speech acts outlined in the preceeding section by introducing the HAEL formalism. We shall discuss the model of atomic propositional speech acts, in which the propositional content is analyzed as an atomic proposition, thereby avoiding the necessity of incorporating a theory of reference into the speech act theory. In subsequent sections, we shall explain how a theory of reference can be integrated with the atomic propositional theory.

An HAEL theory consists of a language $\mathcal{L}'$, together with a set of theories $T_0 \ldots T_n$ and with a partial order $\prec$ among the theories. The base language $\mathcal{L}$ is assumed to contain modal operators $[a]P$ (interpreted as P is true within theory a, or as a believes P if a represents an agent), $[a+b]P$ (interpreted as a and b mutually believe P), and $\{a\}P$, (interpreted as a has goal P.) These operators are assumed to have the properties of a belief logic, such as weak S5. The language $\mathcal{L}'$ is formed by augmenting $\mathcal{L}$ with an additional set of modal operators $L_0 \ldots L_n$. The sentence $L_i P$ is interpreted as P is a theorem of T_i. Conversely, $\neg L_i P$ means P is not a theorem T_i. The structure

of the partial order among these theories is subject to the restriction that, if $\neg L_i P$ is in theory T_j, then $T_i \prec T_j$. In other words, theories cannot express within themselves their own lack of information, but only about such lack relative to theories lower in the hierarchy. This property is important to guarantee the existence of a constructive semantics for theory, and its computational tractability [23].

Appelt and Konolige describe a theory of speech acts with atomic propositional contents based on this formalism. The HAEL theory for atomic propositional speech acts consists of two levels, as illustrated in Figure 6.1. The following axiom describes the content of the utterance theory for utterances of a declarative sentence:

$$[u]P, \quad P \text{ the propositional content of utterance} \tag{2}$$

The following axiom describes the hearer's beliefs as a consequence of the speech act:

$$\text{in } T_1\text{:} \quad ([u]\phi \wedge \neg L_0 \neg [h]\phi \wedge \neg L_0[h]\neg[s]\phi \wedge \neg L_0[h]\neg\{s\}[h]\phi) \supset [h]\phi \tag{3}$$

According to this axiom, the hearer believes ϕ (in T_1) as long as ϕ is derived from the propositional content of the utterance ($[u]\phi$), believing ϕ does not contradict anything from h's general knowledge in T_0, ($\neg L_0 \neg [h]\phi$), it doesn't contradict what he already believes the speaker believes ($\neg L_0[h]\neg[s]\phi$, i.e. the hearer believes that the speaker is not lying), and he doesn't believe anything to contradict the assumption that the speaker is using the utterance with communicative intent ($\neg L_0[h]\neg\{s\}[h]\phi$, i.e., the hearer doesn't believe that the speaker doesn't want him to believe ϕ).

What we must now do is explain how this basic theory of atomic propositional speech acts can be extended to include speech acts that contain referring expressions. In th latter case, the speech act, rather than introducing a proposition ϕ, introduces as its propositional content an individuating set a, of which P is predicated of its referent, and a description D, which is introduced by the referring expression. Before we can do this, we must formally describe the basic elements of our referring model in terms of the HAEL theory.

6.3 Representation of Individuating Sets

Earlier we introduced the *individuating set* as an agent's representation of individual concepts. An individuating set is composed of *descriptions* that pick out the object in the world represented by the individuating set. In our model, individuating sets are represented as a logical theory associated with an agent, and an internal designator. These descriptions play the role of modes of presentation and constitute the axioms of the individuating-set theory. This theory is intended to capture all the information used by the agent to distinguish this individual from others in the world under a particular collection of modes of presentation. The function refis(a, s) maps an agent a and an individuating set s into the referent of that individuating set for that agent. The formula

$$[\text{is}(a, n)]P(\text{refis}(a, n)) \tag{4}$$

represents the fact that P is an individuating property of the individual that is identified by the individuating set n of agent a.

This formal model of individuating sets embodies one of the most important aspects of our theory of individuation: an individuating representation denotes its referent by virtue of its descriptive content. The theory individuates the individual refis(a, n) relative to a model, in the sense that,

given a universe of individuals for interpretation, if the interpretation of all predicates and other constants is fixed, there is only one possible interpretation of refis(a, n).

Individuating properties are really a particular type of belief, so the agent's individuating theories must always be consistent with his beliefs about the object that the theory individuates. This principle is expressed by the following *individuating description belief schema:*

$$[\text{is}(a, x)]\phi \supset [a]\phi. \tag{5}$$

The converse of this axiom does not hold in general, since an agent can have any number of beliefs about an object that are not individuating beliefs. For example A may believe that the person individuated under the property "first president of the United States" chopped down a cherry tree. Individual B may believe that the person individuated by the same property never chopped down any cherry tree. Both individuals have beliefs about the *same* person individuated under identical properties, even though their beliefs about that individual are mutually inconsistent.

One particularly important belief that agents may have about their individuating sets is that two such sets represent the same individual (presumably individuated under a different collection of properties). If this situation obtains for an agent a involving s_1 and s_2, it is stated straightforwardly as

$$[a]\,(\text{refis}(a, s_1) = \text{refis}(a, s_2)) \tag{6}$$

Often, we shall write this as $[a]s_1 \sim s_2$ when there is no ambiguity about whose individuating sets are being refered to.

6.4 Understanding Speech Acts with Referring Expressions

Like the theory of atomic propositional speech acts presented in Section 6.2, the understanding of speech acts that contain referring expressions can also be viewed as a belief revision process. The beliefs about the identity of the individual introduced by the speaker have to be integrated with the hearer's beliefs, as well as what the speaker predicates about it. The model of speech act understanding for speech acts with referring expressions incorporates the model for atomic propositional acts (see Figure 6.1).

Although we are not attempting to develop a theory of discourse here, it is a fact that all speech acts take place as part of some (perhaps very short) discourse, and they take place within a context in which certain individuals may be known *a priori* to be particularly relevant to the discourse. Grosz and Sidner [18] have outlined a theory of discourse based on three components: linguistic structure, intentional structure, and attentional state. The intentional structure consists of a plan shared by the participants (the discourse purpose, or DP), together with beliefs about how a given utterance fits in with subparts of this plan (the discourse segment purpose, or DSP). The attentional state model is a stack-based model of what the participants mutually believe to be individuals who are relevant to a particular DSP. The purpose of the attentional state model is to localize the information that must be considered in order to understand the relationship between an utterance and the DSP, compute how the current DSP changes in response to subsequent utterances, and, which is most important for our model, to constrain the possible referents of referring expressions.

We shall not attempt here to give a full axiomatization of the attentional state theory within HAEL. In fact, it is not clear whether the contribution of attentional state to the discourse model is best represented as a logical theory of relevance. However, for the sake of this analysis, we shall assume that, whatever the details of an attentional state model, it provides certain input to the referring model, and we show how this input is used. One important function of this

attentional state model is to define *what individuals are relevant to the discourse* at any given time. This relevance information is based on those individuals that have been previously and explicitly mentioned in the dialogue, as well as those that are *implicitly* in focus [17] because of their relationship to the explicitly mentioned individuals and the related DSP.

The relevance theory is assumed to provide input to the referring theory by constraining possible coreference relations between the newly recognized individuating set introduced by the utterance and other individuating sets. This constraint, which takes the form of a disjunction of coreference possiblities, is based on focus information, the proposition expressed in the utterance, and deictic gestures, but not on the description used in the referring expression. If, for example, a is an individuating set introduced in the utterance theory, and $c_1 \ldots c_n$ are relevant individuating sets according to this theory, then the following disjunction would be derivable in the relevance theory:

$$a \sim c_1 \vee a \sim c_2 \vee \ldots \vee a \sim c_n \tag{7}$$

Our model is completed with the addition of the *description theory,* which describes how the speaker's referring description is related to the individuating set introduced by the utterance.

Let us assume that a speaker utters a declarative sentence whose propositional content is a monadic predicate of a single argument, to which he refers with a noun phrase. Let us also assume that the speaker and hearer are engaged in an ongoing dialogue about assembling a piece of equipment. The speaker says "You are holding the clear plastic spout." The information entered into the utterance theory from this utterance is

$$[u]\text{Holding}(h, \text{refis}(s, a)) \tag{8}$$

$$[u][\text{is}(s, a)]\text{Clear}(\text{refis}(s, a)) \wedge \text{Plastic}(\text{refis}(s, a)) \wedge \text{Spout}(\text{refis}(s, a)) \tag{9}$$

In this utterance, the speaker represents to the hearer that he has an individuating set for a, containing "clear," "plastic" and "spout" as descriptors.

The speech act theory T_1 of Figure 6.1 contains the atomic propositional speech act belief Axiom 3. This axiom says that whatever is true in the utterance theory gets transferred to the hearer's beliefs at Level 1, provided that it does not contradict anything the hearer believes at Level 0, including his beliefs concerning the speaker's sincerety and communicative intent. If these defaults are not defeated, the speech act belief theory now contains

$$[h]\text{Holding}(h, \text{refis}(h, a)) \tag{10}$$

$$[h][\text{is}(s, a)]\text{Clear}(\text{refis}(s, a)) \wedge \text{Plastic}(\text{refis}(s, a)) \wedge \text{Spout}(\text{refis}(s, a)) \tag{11}$$

We pointed out in Section 5 that the literal goal of referring is for the hearer to have an individuating set representing the same individual as the speaker's individuating set. This literal goal is satisfied by its very recognition. If the hearer recognizes the speaker's intention for him to have an individuating set, then, simply by so doing, he has one. This fact is embodied in the following schema in T_1 called the *activation axiom:*

$$\text{in } T_1\text{:} \qquad [h][\text{is}(s, x)]\phi \supset [\text{is}(h, x)]\text{refis}(s, x) = \text{refis}(h, x) \tag{12}$$

Note that, according to this axiom, the descriptive content of the speaker's individuating set is not included in the hearer's individuating set. Its actual content could be described as "the thing the speaker was talking about when he uttered the referring expression for x."

Let us assume that, according to the relevance theory, the newly introduced individuating set a must be coreferential with one of two other individuating sets, c_1 and c_2. This is expressed as follows:

$$[h]a \sim c_1 \vee a \sim c_2. \tag{13}$$

We have not yet identified the exact contents of the description theory. The purpose of the description theory is to explain how the hearer's beliefs about the referring expression are integrated with his other beliefs, and with his individuating set. In the case of the propositional speech act, the speaker adopts the belief introduced by the utterence unless it conflicts with what he knows, or if he believes the speaker is being insincere. We propose that a hearer adopts the speaker's referring description (and any logical consequence) as part of the individuating description, provided that doing so does not contradict anything in any of the lower theories derived from relevance, or his general knowledge. This leads to the following *description adoption default rule:*

$$\text{In } T_3: \quad ([h][\text{is}(s,x)]\phi \wedge \neg L_2[h]\neg\phi) \supset [\text{is}(h,x)]\phi \tag{14}$$

Assume that in the hearer's belief theory we have

$$[h]\text{spout}(c_1) \wedge \neg\text{spout}(c_2) \tag{15}$$

Given this information, it is perfectly consistent for the hearer to adopt the entire description in his own individuating set for a. Given this adoption of the description, plus his general beliefs, it is then possible to conclude $[h]a \sim c_1$ in the Description Theory.

6.5 Referring and Mutual Belief

As we have presented it so far, the model is inadequate to account for some observations and intuitions about reference. Clark and Marshall [7] illustrate that, for one agent to refer to an individual for another using a description D, it is insufficient for the hearer to merely reason with his private beliefs about what D denotes, but he must also take the speaker's beliefs into consideration. Futhermore, it is possible to construct increasingly complex, but intuitively clear situations in which any arbitrary finite limit on the nesting of the hearer's beliefs regarding the speaker's beliefs about the denotation will fail to pick out the right referent. The model we have presented so far deals only with the speaker's and hearer's individual beliefs.

Perrault and Cohen [32] point out that a slightly weaker condition on mutual belief can suffice, namely, that the speaker and hearer *agree* that the description denotes the referent, with "agree" defined formally as

$$\text{agree}(s,h,\phi) \equiv_{def} [s+h]\phi \vee [s][s+h]\phi \vee [s][h][s+h]\phi \vee \ldots, \tag{16}$$

or, in other words, that only some finite number of instances of the mutual belief schema does not hold in the given situation. This weakened version of the mutual belief requirement enables Cohen and Perrault to explain how reference succeeds in a situation like the following:

> S and H are at a party. They watch together as water and gin are being poured in two identical glasses and given to women W_1 and W_2 respectively. Later S sees H see the women swap glasses, without seeing H see him. S also overhears A telling H that S saw him see the exchange. Later, S tells H: "The woman with the martini is the mayor's daughter."

Here it seems that S has successfully referred to W_2, even though S believes W_2 is drinking water, S believes that H believes that, and even that S believes that H believes that S believes W_2 is drinking water. The reason is that the identification of the referent is based on what the two agents agree upon, rather than on any one of their private beliefs.

Even the relatively weaker agreement condition of Cohen and Perrault poses some difficult problems when stretched to account for a wide range of referring actions. For example, consider the problem of accounting for the *subsumption* of an informing action by a referring expression [4]. Often a speaker can augment a description with an additional descriptor, with the intention that the descriptor function not to pick out the referent, but to *inform* the hearer that the referent has the property indicated by the descriptor. The speaker, of course, assumes that the hearer can understand the reference based on contextual cues and the remainder of the description. For example, a speaker might give the instruction "Remove the casing with the 3/8-inch wrench in the toolbox in the cabinet." The speaker knows that there is only one 3/8-inch wrench, and that the descriptor is sufficient for the hearer to know what he is talking about. However, he knows that, to carry out the request, he has to get the wrench and, to do that, he must know its location. Therefore, the location is included as part of the description, which serves both to refer and inform.

If the agreement condition is taken seriously, the speaker and hearer could never agree on the denotation of the description because the speaker and hearer do not agree at *any* level that the part of the description constituting the informing descriptor holds of the referent. If such knowledge existed, there would be little point in the speaker's performance of the informing action in the first place.

Cohen and Perrault observe a similar problem with respect to attributive references. In the paradigmatic attributive cases, there is no agreement (in the technical sense) to the effect that only one contextually relevant individual exists that could satisfy the description. This observation leads them to conclude that referential and attributive referring expressions constitute the performance of two different types of acts because their preconditions are different.

The problem faced by these accounts of mutual belief and reference is not that the condition of mutual belief is too strong (as the ability to construct examples points out), but that the conditions under which mutual belief or agreement can be said to hold are too strong. Consider the case in which A and B attend a movie M, A sees B in the audience, but B does not see A. The next day A approaches B and says "Did you enjoy the movie last night?" In this case A and B have no mutual knowledge (not even agreement) about any particular movie. However, in the typical case B is unlikely to spend much time agonizing over what A means, particularly if he knows from his own private beliefs that it couldn't be anything other than M that A is talking about. He may indeed wonder how it is that A came to know about M, but *as long as an assumption of agreement doesn't conflict with anything else he knows,* he can assume that A is talking about M.

A propositional model of referring based on nomonotonic reasoning about mutual belief was first proposed by Nadathur and Joshi [30]. Their model was based on a logic devised by Konolige [22] whereby a provability operator is introduced into a propositional modal logic. They suggested that the prerequisites enabling a speaker s to use ϕ as a referring description for hearer h are

$$[s][h]\phi, \quad \text{and it is consistent to believe} \quad [s+h]\phi. \tag{17}$$

The problem with this attempt is that the theory makes no prediction about what happens to the hearer's belief when niether condiditon is satisfied; it therefore cannot give an adequate account of the Cohen and Perrault martini examples, such as the one cited here. It is merely offered as a weaker alternative to mutual belief.

In the nonmonotonic model we are proposing, the hearer, as in Nadathur and Joshi's model, can use his private knowledge to determine the referent, but this conclusion can be allowed only if

there is no explicitly present belief that would contradict the assumption that the description could be mutually believed. In addition, when a conclusion about the referent based on private belief is defeated, we must somehow explain within the theory the hearer's ability to back off from his own beliefs and reason instead about his beliefs about the speaker's beliefs, etc.

Fortunately, the extensions that must be made in the referring model to account for mutual belief are relatively small. First, we claim that the description introduced by the utterance must entail consequences of mutual belief. The speaker, when using a referring description ϕ, in effect asserts (subject to contradiction by evidence, of course) that "ϕ is true of the thing I am referring to, and I believe that this fact is mutually believed." This can be represented by axiomatizing the initial utterance theory as follows:

$$[u][\text{is}(s,a)](\phi \wedge [s][s+h]\phi) \tag{18}$$

Then the description adoption default rule (14) is replaced by the following schematic generalization:

$$\begin{aligned} &\text{In } T_3: \\ &[h][\text{is}(s,x)]\phi \wedge \neg L_2\,(\forall y\ ([h]\phi \supset [h]x \sim y) \supset [h]\neg[s+h]\phi(y|x)) \\ &\quad \supset [\text{is}(h,a)]\phi \end{aligned} \tag{19}$$

The schema (19) is complex but can be glossed fairly straightforwardly. It says that, if ϕ is a consequence of the individuating theory introduced by the utterance, then h will transfer ϕ to his own individuating set *unless* it is provable in the relevance theory that, if h believed ϕ and doing so constrains the possible referents of the individuating set to some individual y, then h believes it is not mutually believed that ϕ holds of y. The notation $\phi(y|x)$ means that y is substituted for x wherever it occurs in ϕ.

Let us consider how the Perrault and Cohen example would be handled by this model. Let $M(x)$ mean "x is drinking a martini" and $W(x)$ mean "x is drinking water." Suppose S says to H "The woman with the martini is the mayor's daughter."
Utterance theory:

$$[u]\text{mayor's-daughter}(a) \tag{20}$$

$$[u]\,([\text{is}(s,a)]M(a) \wedge [s][s+h]M(a)) \tag{21}$$

General knowledge in T_0:

$$[s+h]\forall x W(x) \equiv \neg M(x) \tag{22}$$

$$[h]W(w_2) \wedge M(w_1) \tag{23}$$

$$[h][s]W(w_2) \wedge M(w_1) \tag{24}$$

$$[h][s][h]W(w_2) \wedge M(w_1) \tag{25}$$

$$[h][s][h][s]W(w_1) \wedge M(w_2) \tag{26}$$

Relevance in T_2:

$$[s+h]a \sim w_1 \vee a \sim w_2 \tag{27}$$

We attempt to derive in T_3 what the contents of H's individuating set are. Let us attempt to prove that $[h][\text{is}(h,a)]M(a)$. According to Axiom 19, this will hold if $[h][\text{is}(s,a)]M(a)$ (it will, if the literal goal was recognized), and it is not a theorem of T_2 that

$$\forall y\ ([h]M(a) \supset [h]a \sim y) \supset [h]\neg[s+h]M(w_1). \tag{28}$$

We note that the axioms in T_0 combined with T_2 imply that, if $[h]M(a)$, then $a \sim w_1$. We observe that the consequent of (28) matches a consequent of (26), namely, $[h]\neg[s][h][s]M(w_1)$. Therefore (28) *is* a theorm of T_2, and the default that transfers the belief $M(a)$ to H's individuating set is defeated. A similar analysis applies when $M(a)$ is replaced by $[s]M(a)$ or $[s][h]M(a)$. However, the only general knowledge for determining the identity of a based on $[s][h][s][h]M(a)$ is (26). Since there are no other beliefs about mutual belief to allow (28) to be derived, h's individuating set contains $[s][h][s][h]M(a)$ and, according to the relevance theory, a must be coreferential with w_2.

What is interesting about this example is that it demonstrates that it is unnecessary to postulate explicit mutual knowledge of (or agreement on) a referring description's denotation in order to demonstrate that a speaker and hearer can concur about what is being talked about, given constraints stemming from their private knowledge and a discourse model. Cohen and Perrault assert that this precondition of mutual belief is an essential distinction between the referential and attributive uses of referring expressions, leading them to conclude that referring referentially and referring attributively are fundamentally different actions.

The fact that the mutual belief or agreement precondition can be weakened when default reasoning is employed raises the interesting possibility that a single referring model can account adequately for both referential and attributive uses. We believe this hypothesis is correct and is supported by the following discussion.

7 The Referential-Attributive Distinction

In this section, we examine the manner in which our model can represent the famous distinction between referential and attributive uses of definite descriptions (Donnallen's distinction, for short [12]). Donnellan's distinction is an excellent test case for any theory of reference, particularly for a model like ours that is based on the descriptive approach. After all, Donnellan's distinction has been used extensively in many arguments against the descriptive program [28,29,27].

Let us recall, then, the two crucial features of Donnellan's distinction: whereas, in the attributive use, the description must be satisfied for reference to succeed, in the referential use this is not so. In the latter case, moreover, the speaker has a particular object in mind, while in the attributive case he does not.

How, then, should Donnellan's distinction be interpreted? As Kronfeld has shown [24,27], the distinction has three aspects — the epistemic, the modal, and the denotational, that should be treated separately.[8] The epistemic aspect has to do with the question of whether, in using a definite description, the speaker in some sense knows who or what the referent is. Typically (but not necessarily always), in the referential use he does know, while in the attributive use he does not. The modal aspect has to do with the question of whether the definite description is used as a rigid designator. In the referential use it is so employed, while in the attributive use it is not. Finally, the denotational aspect has to do with the question of whether the definite description must denote one and only one object for the speech act to be successful. Typically, in the attributive use it does, while in the referential use it does not. Table 1 summarizes these three aspects.

The three aspects of Donnellan's distinction are conceptually independent of one another. Yet they must all be bound together in an obvious way, or else Donnellan's distinction would not have

[8] In Kronfeld's earlier paper [24], the denotational aspect of Donnellan's distinction was called "the speech act aspect."

such a persuasive ring to it. What needs to be explained, therefore, is the intuitive immediacy of the distinction, given its complexity. How is it that there is in fact a single referential/attributive distinction and not three?

The answer, no doubt, must lie in the way the three aspects interact with one another. In the paradigmatic examples of attributive usage, the speaker selects a *particular* mode of presentation (the modal aspect) as well as, not surprisingly, a definite noun phrase expressing that presentation mode (hence the denotational aspect). As Donnellan himself points out [12], *lack* of knowledge concerning the referent (the epistemic aspect) is not a prerequisite of attributive usage. One can know who murdered Smith, yet insist that *anyone* who would have murdered Smith in such a brutal way must be insane. Nevertheless, the attributive use typically occurs when the identity of the referent is unknown, because, under such circumstances, the intention to refer to an object *as* the such-and-such is easier to recognize. When the identity of the referent is well known to all participants in the conversation, the speaker must frequently make his intention to refer to the object *as* having a certain property more explicit. Moreover, when the speaker has no knowledge of the referent whatsoever, his use of a definite description is not likely to be intended as a rigid designator. We can thus see why the attributive side of Donnellan's distinction encompasses a natural category of uses of definite descriptions. Nonrigid designation tends to accompany a particular choice of a definite description, usually when knowledge of the referent is lacking.

In the paradigmatic examples of referential usage, on the other hand, knowledge of the referent must be possessed; at the same time, the speaker does not consider any *particular* mode of presentation to be of importance. The speaker's goal is simply to have the hearer identify appropriately what is being talked about, a task for which any well-suited referring expression would be acceptable. Moreover, when knowledge (including mutual knowledge) of the referent exists, identification can be achieved, to a large extent, apart from the descriptive content of the referring expression. Hence, strict denotation is frequently unnecessary.

Let us see now how all this can be expressed in our model. The first aspect of the distinction, the *epistemic* one, centers on the question of how much and what kind of knowledge the speaker (and hearer) posess concerning the referent. Along this dimension, a referring expression is referential only if the speaker knows some particular individuating property of the referent that is relevant to the task at hand, otherwise it is attributive. Note that the characterization is done purely in terms of internal states, that is, in terms of the descriptive content of individuating sets.

The denotational aspect of Donnellan's distinction has been utilized extensively in arguments against the descriptive approach. In a nutshell, according to a descriptive theory, reference is a function of descriptive content. But then, it is argued, if this is true, how could a speaker refer successfully while using erroneous descriptions?

Conceptually, the problem is solved by insisting that the descriptive approach provides a research program for a theory of mind, not a theory of the meaning of singular terms [27]. In terms of our model, the problem is solved by showing how the literal goal of referring can be achieved and the

Donnellan's Distinction					
Mental-State Criterion				Denotation Criterion	
Epistemic Aspect		Modal Aspect		Denotational Aspect	
Referential	Attributive	Referential	Attributive	Referential	Attributive
Knowledge of the referent	No knowledge of the referent	Rigid designation	Nonrigid designation	Reference without denotation	No reference without denotation

Table 1: Aspects of Donnellan's distinction.

identification constraints satisfied regardless of whether the descriptive content of the noun phrase denotes nothing, is incorrectly applied, or is lacking altogether. Furthermore, the default-based analysis shows how the description used to identify the referent does not have to be identical to the description used in the utterance. The description adoption default rule (14) provides a means of combining, in a consistent manner, information from the description with other information from the discourse and the agent's general beliefs. The result is that an individuating set for the referent is obtained, and the descriptive content of that set determines the referent. If the description is inaccurate, its content is weakened in a systematic way to make it consistent with all the other sources of information that bear on the agent's belief revision process. An example may illustrate this point.

Suppose that the set of physical objects relevant to the current discourse consists of a pink ball and a green block (represented by the hearer's individuating theories b_1 and b_2, respectively). Suppose, furthermore, that all of the following axioms hold in every theory (including individuating theories):

$$\forall x \, \mathrm{red}(x) \supset \neg \mathrm{pink}(x) \tag{29}$$

$$\forall x \, \mathrm{ball}(x) \supset \neg \mathrm{block}(x) \tag{30}$$

Suppose the speaker says "Pick up the red ball." The initial contents of the utterance theory are $\{s\}\text{pick-up}(h, \text{refis}(s, rb))$ and $[is(s, rb)]\text{red}(\text{refis}(s, rb)) \wedge \text{ball}(\text{refis}(s, rb))$. Let us assume that, in the hearer's relevance theory, it is possible to derive $[h]b_1 \sim rb \vee b_2 \sim rb$. In the description theory, the description adoption default asserts that the hearer will accept any part of the speaker's description, as long as the result is consistent with what the hearer believes. Therefore, $[\text{is}(h, rb)]\text{red}(\text{refis}(h, rb)) \wedge \text{ball}(\text{refis}(h, rb))$ is inconsistent with the relevance theory (because no relevant thing is a red ball), but another weaker consequence of the theory, $[\text{is}(h, rb)]\text{ball}(\text{refis}(h, rb))$ is consistent, provided that $[h]b_1 \sim rb$. If the individuating theory for b_1 contains enough descriptive information to allow the hearer to carry out the request and pick it up (e.g. $[is(h, b_1)]$ contains information about, say, its position and orientation), then the hearer has identified the referent as well.

Goodman [13] analyzed a number of task-oriented dialogues in which, frequently, the speaker succeeds in referring by using inaccurate descriptions. He describes a process of *relaxation* by which a weaker version of a description is adopted by the speaker than the one actually used. The nonmonotonic model formalizes this idea of relaxation. Naturally, to avoid having competing inconsistent theories, some strategy of selecting an ordering on the predicates to be weakened must be adopted. A discussion of this belief revision topic is beyond the scope of this article, but the interested reader is directed to Goodman's article for a justification of one relaxation strategy.

It is interesting to note that the mutual-belief example of the preceeding section is an instance of exactly the same mechanism. In the martini example, the description used by the speaker is weakened to successively deeper instances of the mutual-belief schema until, finally, the description in the hearer's individuating set is consistent with all his beliefs about their shared belief.

This mechanism explains why referring is not entirely dependent on successful denotation. Once this is understood, however, the converse question arises. Why is it that in the paradigmatically attributive use the description *must* denote the referent? If descriptive content is not necessary for successful reference, what is so special about the attributive use? Why is the descriptive content in such cases *essential*? A partial answer to that question has to do with relative contributions of different theories toward identification. If the relevance theory furnishes no clues, the utterance theory must be relied upon, which means that the descriptive content of the referring expression is

the major source of identifying information. This is what happens when very little is known about the referent. In such cases, no facts known to the hearer can block incorporation of the descriptive content of the referring expression into an individuating set.

However, this account of the attributive side of the denotational aspect cannot be the entire story. As Donnellan himself points out, the identity of the referent can be established independently of the description used; nevertheless denotation is still necessary for the speech act to succeed. Focusing on the modal aspect of Donnellan's distinction provides an explanation. Kronfeld [24,27] suggests an interpretation of the modal aspect that is based on two types of referring intentions. In the first type, the role of the referring expression is simply to identify the referent efficiently. In the second type, a particular description (called the *conversationally relevant description*) is used to convey a certain implicature. Compare, for example, the following two sentences:

7.1 *Jones has to be insane.*

7.2 *The man who killed Smith in such a brutal way has to be insane.*

In the second utterance, it is implicated that the main reason for the belief that the murderer is insane is the very fact that he committed such a brutal homicide. Since the descriptive content plays a role in interpreting the speech act as a whole, failure of denotation may cause the speech act to be incoherent even though reference may be secured. For example, if sentence 7.2 is used to refer to Jones at the dock, the hearer may very well understand whom the speaker is talking about. But, if Jones is proved innocent because Smith's death was a bizarre suicide, the speaker would have to withdraw his assessment. The description he employed is *conversationally* relevant, and therefore failure of denotation contaminates the speech act as a whole.

An adequate computational treatment of the modal aspect, however, must be part of an adequate computational theory of conversational implicatures. This is obviously beyond the scope of this paper. In any case, it should be clear that both the "referential" and the "attributive" can be accounted for by a single, uniform, referring mechanism.

8 Conclusion

In this article we have argued that, to provide a computationally useful model of referring, one of the central questions to be considered is how it is that mental representations of objects are related to the things they stand for. After consideration of the alternatives, it appears that the most promising strategy is to assume that mental representations of individuals incorporate descriptive content, and that it is this content that determines what individual the representation represents. We have therefore proposed a structure called an *individuating set* as containing precisely the information that defines the essential properties of an individual from some conceptual perspective.

By using the individuating set as a tool, it becomes possible to state precisely what it is that a speaker wishes to accomplish with a referring act in terms of changes in the hearer's beliefs and his individuating sets. When a speaker refers to something, he intends that the hearer acquire an individuating set representing the same individual as one of his own individuating sets. This is the *literal goal* of a referring act and, as with other Gricean intentions, it can be satisfied merely by the hearer's recognition of the speaker's intention. In addition, the hearer must recognize whether he must posess a certain kind of knowledge of the referent to have understood the speaker. This knowledge is called *identification constraints*, and satisfying the relevant identification constraints is the *discourse purpose* of the referring act.

Many philosophical arguments have been advanced against the descriptive paradigm, many of them centered on the referential-attributive distinction. The central question is, "if reference

is determined by descriptive content, then how can a speaker succeed in referring by using an inaccurate description?" Furthermore, there seems to be a class of referring acts in which the verity of the description is essential to success of the speech act (attributive use) whereas in others (referential use) the speech act can succeed even when the description is true of nothing, or is true of an unintended individual.

We have shown that, by constructing a logical model of referring based on a nonmonotonic logic, it is possible to explain differences between the description used by the speaker and changes in the hearer's mental state that result in individuating the referent under a different set of properties than those used by the speaker. The basic intuition is that the speaker's description is incorporated into the hearer's individuating set, as long as doing so is consistent with his other beliefs, as well as his beliefs about their mutual beliefs. This model of referring accounts for the success of different kinds of discrepancies between the speaker's and hearer's beliefs that arise in referring, including the inaccurate descriptions studied by Goodman, based on the physical properties of the individuals referred to, and the mutual-belief related discrepancies studied by Cohen and Perrault.

The nonmonotonic model provides a uniform account of both the referential and attributive uses of referring expressions, which had previously been thought to be different action types because a mutual belief precondition that is present in the referential case is absent in the attributive case. The "essentiality" of the descriptive content in typical attributive uses can be explained by the hearer's lack of specific knowledge of the referent that could block defaults, and by the fact that the speaker can intend a conversational implicature based on the propositional content of the referring expression, which makes the specific propositional content of the referring expression essential to the success of the speech act as a whole.

This work raises a number of questions that demand further research. One important question is how identification constraints are recognized. It is fairly simple to state identification constraints in task-oriented dialogues, which frequently involve perception and manipualtion of the objects involved. However, when one considers more general domains, it is not always so obvious how such constraints can be determined. The determination of the discourse segment purpose is essential to determining the discourse purpose of the referring act.

The nonmonotonic model still presents some serious problems with regard to control of inference if it is incorporated into a computational system. The specific theory of the domain must be decidable in order to guarantee the decidability of the nonmonotonic theory. If developing such a decidable theory is impossible, then decidability must be approximated with a strategy based on resource-bounded computation.

The nonmonotonic theory declaratively describes a relationship between propositions in the utterance theory and the speaker's and hearer's beliefs. If the transfer of a specific belief from the utterance theory is blocked by conflicting evidence, then a weaker consequence of that theory can be transfered. The number of consequences of a theory is infinite — obviously not all can be considered individually. Therefore, some ordering strategy for consideration of weak consequences must be implemented. Goodman [13] offered one such strategy, however, it remains to be determined how such a strategy would interact with the strategy of considering deeper nestings of shared belief, as in the Cohen and Perrault examples. Also, different ordering strategies could result in different determinations for the referent of a description.

In spite of these challenges, the nonmonotonic descriptive model provides a very promising foundation for research on a computational model of referring.

References

[1] James F. Allen. *Recognizing Intention in Dialogue.* PhD thesis, University of Toronto, 1978.

[2] James F. Allen and C. Raymond Perrault. Analyzing intention in dialogues. *Artificial Intelligence*, 15(3):143–178, 1980.

[3] James F. Allen and C. Raymond Perrault. Participating in dialogues: understanding via plan deduction. In *Proceedings, Canadian Society for Computational Studies of Intelligence*, 1978.

[4] Douglas E. Appelt. *Planning English Sentences.* Cambridge University Press, Cambridge, England, 1985.

[5] Douglas E. Appelt. A practical nonmonotonic theory for reasoning about speech acts. In *Proceedings of the 26th Annual Meeting*, Association for Computational Linguistics, 1988.

[6] Robert C. Moore Semantical Considerations on Nonmonotonic Logic. *Artificial Intelligence*, 25(1), 1985.

[7] Herbert Clark and C. Marshall. Definite reference and mutual knowledge. In A. Joshi, I. Sag, and B. Webber, editors, *Elements of Discourse Understanding*, Cambridge University Press, Cambridge, England, 1978.

[8] Philip R. Cohen. *On Knowning What to Say: Planning Speech Acts.* PhD thesis, University of Toronto, 1978.

[9] Philip R. Cohen. The need for identification as planned action. In *Proceedings of the Seventh Annual Conference on Artificial Intelligence*, pages 31–36, 1981.

[10] Philip R. Cohen and H. Levesque. Speech acts and rationality. In *Proceedings of the 23rd Annual Meeting*, pages 49–59, Association for Computational Linguistics, 1985.

[11] Philip R. Cohen and C. Raymond Perrault. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3:117–212, 1979.

[12] Keith S. Donnellan. Reference and definite description. *Philosophical Review*, 75:281–304, 1966.

[13] Brad Goodman. Reference identification and reference identification failures. *Computational Linguistics*, 12(4):257–272, 1986.

[14] H. Paul Grice. Utterer's meaning, sentence meaning, and word meaning. *Foundations of Language*, 4:225–242, 1968.

[15] H. Paul Grice. Utterer's meaning and intentions. *Philosphical Review*, 78:147–177, 1969.

[16] H. Paul Grice. Meaning. *Philosophical Review*, LXVI(3):377–388, 1957.

[17] Barbara J. Grosz. *The Representation and Use of Focus in Dialogue Understanding.* Technical Report 151, SRI International Artificial Intelligence Center, 1977.

[18] Barbara J. Grosz and Candace L. Sidner. Attentions, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, July-September 1986.

[19] Hans Kamp. A theory of truth and semantic representation. In Groenendijk et. al., editor, *Truth, Interpretation, and Information*, Foris, Dordrecht, Netherlands, 1984.

[20] Asa Kasher. Conversational maxims and rationality. In A. Kasher, editor, *Language in Focus*, pages 197–216, D. Reidel Publising Co., Dordrecht-Holand, 1976.

[21] Asa Kasher. Gricean inference revisited. *Philosophica*, 29(1):25–44, 1982.

[22] Kurt Konolige. Curcumscriptive ignorance. In *Proceedings of AAAI-82*, pages 202–204, 1982.

[23] Kurt Konolige. A hierarchic autoepistemic logic. 1988. forthcoming technical note.

[24] Amichai Kronfeld. Donnellan's distinction and a computational model of reference. In *Proceedings of the 24th Annual Meeting*, pages 186–191, Association for Computational Linguistics, 1986.

[25] Amichai Kronfeld. *Donnellan's distinction as an adequacy test for a referring model.* Technical Note, Artificial Intelligence Center, SRI International, 1988.

[26] Amichai Kronfeld. *Reference and Denotation: The Descriptive Model.* Technical Note 368, SRI International Artificial Intelligence Center, October 1985.

[27] Amichai Kronfeld. *Reference and computation: an essay in applied philosophy of language.* Cambridge University Press, Cambridge, England, Forthcoming.

[28] Amichai Kronfeld. *The Referential Attributive Distinction and the Conceptual-Descriptive Theory of Reference.* PhD thesis, University of California, Berkeley, 1981.

[29] Amichai Kronfeld. *The descriptive approach to reference: why it is difficult to live with, and why we have to.* Technical Note, Artificial Intelligence Center, SRI International, 1988.

[30] Gopalan Nadathur and Aravind Joshi. Mutual beliefs in conversational systems: their role in referring expressions. In *Proceedings of IJCAI-83*, pages 603–605, 1983.

[31] C. Raymond Perrault. *An Application of Default Logic to Speech Act Theory.* Research Report CSLI-87-90, 1987.

[32] C. Raymond Perrault and Philip R. Cohen. Inaccurate reference. In A. Joshi, editor, *Formalizing Discourse*, Cambridge University Press, Cambridge, England, 1980.

[33] James F. Allen Perrault, C. Raymond and Philip R. Cohen. Speech acts as a basis for understanding dialogue coherence. In *Theoretical Issues in Natural Language Processing-2*, pages 125–132, University of Illinois at Urbana-Champaign, 1978.

[34] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13, 1980.

[35] Stephen Schiffer. The basis of reference. *Erkenntnis*, 13:171–206, 1978.

[36] John Searle. *Speech Acts: An Essay in the Philosophy of Language.* Cambridge University Press, Cambridge, England, 1969.

[37] Candace L. Sidner. Plan parsing for intended response recognition in discourse. *Computational Intelligence*, 1(1):1–10, 1985.

[38] Candace L. Sidner. What the speake means: the recognition of speakers' plans in discourse. *International Journal of Computers and Mathematics*, 9(1):71–82, 1983.

[39] Bonnie L. Webber. So what can we talk about now? In M. Brady and R. Berwick, editors, *Computational Models of Discourse*, pages 331–371, MIT Press, Cambridge, Massachusetts, 1983.