SRI International

A COMPUTATIONAL MODEL OF REFERRING

Technical Note 409

January 13, 1987

By: Douglas E. Appelt and Amichai Kronfeld

Artificial Intelligence Center
Computer and Information Sciences Division

APPROVED FOR PUBLIC RELEASE:
DISTRIBUTION UNLIMITED

This research has been made possible in part by the National Science Foundation under Contract DCR-8407238, SRI Project 7985.

SRI International®

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

ABSTRACT

In this paper we present a theory of referring. This theory is presented within the framework of a general theory of speech acts and rationality advanced by Cohen and Levesque. Understanding a speech act involves two tasks on the part of the hearer: recognition of the *literal goal*, and recognition and satisfaction of *identification constraints*. We present a theory in which it is possible to demonstrate that, if a referring expression is uttered under appropriate circumstances, the literal goal is achieved. Furthermore, the same theory can account for the fact that referring expressions can be used under other circumstances to inform and make requests. This theory has application to the design of a natural-language utterance planning system that formulates utterances to achieve its goals.

1 Motivation and Preliminary Discussion

When two agents talk to each other, they had better both know what they are talking about. This, we hasten to add, is not a condescending directive for nonexperts to keep quiet. It is rather a general rationality constraint on the production of speech acts. Speech acts frequently mention *things:* one may promise to return a book, request that a window be closed, assert that a certain person is missing, and so on. But two rational agents can expect such speech acts to be successful only if both know precisely what book, window, or person is being discussed. Thus, the problem arises: *how do speaker and hearer form an agreement as to which entities are being talked about?*

1.1 What is Referring?

Before we can present a theory of referring, it is necessary to specify what referring is. This is a question that has occupied several generations of philosphers of language, nor is it likely that anything close to a consensus will emerge in the forseeable future. For example, Donnellan [6] has argued that using definite descriptions attributively does not constitute referring at all, since in such cases the speaker "does not have any particular object in mind." For Searle [15], on the other hand, there is no significant difference between the attributive and the referential use, as far as referring is concerned. It would be impractical for us to wait until the philosphical problem of reference is resolved. Instead we select a philosophical position that appears promising as a foundation for our computational model. No harm is done as long as we remain aware of our philosophical prejudices. We therefore define referring from an intuitive standpoint as follows:

> An agent is referring when he has a mental representation of what he believes to be a particular object, and he intends the hearer to come to have a mental representation of the same object, at least in part through the use of a noun phrase that is intended to be a linguistic representation of that object.

Note that, according to this definition, indefinite descriptions could function as referring expressions.

1.2 Internal and External Perspectives

In dealing with the problem of interpreting noun phrases, two perspectives may be distinguished. One perspective is *internal,* the other *external* with respect to the discourse. According to the internal perspective, theorists seek to provide constraints on the relations among noun phrases, and they need not concern themselves with the question of determining what objects (if any) correspond to these noun phrases. For example, a major theoretical task associated with the internal perspective is determining constraints under which a pronoun or a definite description may be anaphorically linked to other noun phrases. Within this framework, it makes little sense to distinguish betweeen, say, "the student," and "the average student," as far as their potential for

binding pronouns is concerned. The question of whether there is a *particular* student in the real world who is supposed to be identified by the hearer need not have much interest for a theorist adopting the internal perspective. From the *external* perspective, on the other hand, the relation (if there is any) between a particular object and a noun phrase is of prime importance.

The internal perspective is essential to understanding the syntax and semantics of natural language. However, once we consider language as a way for rational agents to achieve goals, the external perspective becomes indispensable. After all, it makes a great deal of difference to an agent who is attempting to achieve some goal through the use of language, to know whether the hearer understands that he (the speaker) has a particular object in mind, and furthermore, whether the hearer is able to identify it. Any system that combines linguistic and nonlinguistic actions, and that is capable of cooperative behavior, must be able to talk about objects. It must understand when a noun phrase has a referent in the real world and when it does not, when knowledge of the referent is required and when it is not, when knowledge of the referent is presupposed and when it should be actively sought.

1.3 Objectives of this Analysis

As a study in artificial intelligence, the ultimate goal of this research is the construction of an intelligent agent that uses and understands natural language effectively. Of course, the effective planning and use of natural language sentences requires the effective planning and use of referring expressions. In this paper we provide a *competence* theory of referring, i.e., we provide a theory that states how a referring expression is related to the beliefs and goals of the participants in a discourse. As such it is a theory about agents — it constrains how they can be designed, but it is not a design per se. In the KAMP utterance planning system [2], Appelt used axioms and a theorem prover to model the mental process of an agent planning utterances. The possibility of constructing such a system based on axioms of referring is an important motivation for this work.

Agents do not perform referring actions for their own sake. Referring is always subordinate to some other goals the speaker may have concerning what the hearer should believe or do. Therefore, a theory of reference has to be incorporated into a larger theory of speech acts and rational action. Cohen and Levesque [5] have developed a speech act theory that is ideally suited to this enterprise. A major objective of their research was to characterize illocutionary acts in terms of what the speaker and hearer mutually believe about each other's mental states as a result of the action. We adopt the same objective with respect to the act of referring.

Cohen and Levesque are able to demonstrate how an utterance of a declarative sentence, for example, can be taken as an assertion under one set of circumstances, a promise under another set of circumstances, or a request under still other circumstances. Similarly, we intend to show how the utterance of a referring expression can be used to achieve different kinds of goals, depending

on its context of use.

In certain contexts, referring expressions have the effects of informing and requesting. Appelt [2] observed that one way an utterance may satisfy multiple goals is by using a referring expression to inform the thearer that some property holds of the intended referent. Thus, if a speaker points to a tool on the table, and says "Use the wheelpuller to remove the flywheel," the speaker can rely upon the pointing action to enable the hearer to identify the referent of "the wheelpuller," while the descriptive content of the referring expression serves to inform him that the tool is a wheelpuller.

Appelt's KAMP system was capable of planning referring expressions to satisfy multiple goals. However, according to the theory embodied in KAMP, the capability for referring expressions to inform the hearer derived directly from the presupposition that the speaker believes his description to be true of the referent. The analysis did not rest on Gricean intended recognition of intention, and therefore, informing via a referring expression had fundamentally different effects from informing via a declarative sentence — a distinction we now feel is not well motivated empirically. In this paper we intend to show that given the right theory of reference, the ability to inform using a referring expression follows directly from it.

Cohen [4] observes that in task oriented dialogues, hearers would respond to referring expressions as though they were requests. They would search the area in front of them to determine what the speaker was talking about, and confirm or deny their ability to find the referent even before the speaker could finish his utterance. Cohen concludes that referring is a specific instance of the more general action of making a request, and this made the hypothesis of a separate category of propositional acts of referring unnecessary. While we do not agree that all instances of referring are actually instances of requesting, the *possibility* of using referring as requesting should follow from a proper theory of reference. One of our objectives in this papers is to show that this conclusion follows from a proper theory of reference as well.

A correct analysis of referring must also explain the role of the descriptive content of referring expressions. It is certainly possible for a hearer to figure out the correct referent even though the speaker is wrong and his description is not true of the object he has in mind. A familiar example is the case where the description "the man over there drinking a martini" is used while as a matter of fact the man is drinking water. A hearer may still be able to identify the right individual, and therefore, the referring act may be successful even though the description was wrong.

Finally, one of our objective in this paper is to eliminate what we call the *standard name assumption*. The simplest way to handle reference in language systems is to assume that all objects have standard names. If, for example, a screwdriver is one of the objects in the domain of discourse, then under the standard name assumption, it is represented by some constant, say S_1, which serves as a standard name for the tool. If a speaker says "The screwdriver is broken," referring to the object named S_1, and the hearer believes "the object named S_1 is broken," then and only then is the referring act successful. This is how referring has been implemented in most natural language

systems, including Appelt's KAMP [2].

There are, however, several difficulties with this approach. First, the assumption that all objects have standard names is not plausible in complex situations where there are many objects, some of which are not known to one of the agents. Second, the standard-name approach has some epistemological consequences that are much too strong. For example, if a speaker has successfully referred to the same object on two different occasions, then, under the standard-name assumption, it follows necessarily that the hearer *knows* it was the same object each time. But this is obviously too strong a requirement for successful reference. Suppose that on two separate occasions a speaker refers to a screwdriver, which the hearer picks up and uses for a particular task. Clearly, the referring action is successful both times by any reasonable criterion, and the hearer's knowledge that he has actually manipulated the very same object on both occasions is irrelevant. As will be shown later on, in our model it is possible to discard the standard name assumption altogether.

Our task, therefore, is to develop a theory of referring with five adequacy criteria in mind. In particular, the theory should demonstrate:

1. How a suitably formulated version of the definition of referring given above follows from the speaker's use of a noun phrase under the appropriate conditions.
2. How under the appropriate conditions, a referring expression has the same effect on the hearer as informing him that the referent satisfies the descriptive content of the referring expression.
3. How under the appropriate conditions, a referring expression has the same effect on the hearer as a request that he perform an action mutually believed to be necessary for referent identification.
4. How a referring act can succeed even though the hearer believes the description used is false of the speaker's intended referent.
5. How the standard name assumption can be eliminated.

2 The Goals of Referring Acts

The view that the referring act is a planned effort to achieve certain goals by linguistic means follows from the fact that referring is a *speech act*: all speech acts are attempts to achieve certain goals by linguistic means. However, referring acts differ significantly from illocutionary acts such as asserting and requesting, along two different dimensions which we describe in turn.

Literal goals. In performing a speech act, a speaker may have many distinct goals. For example, by saying "The house is on fire!", a speaker may have the goals of informing the hearer that the house is on fire, frightening the hearer half to death, and making the hearer leave. Only the first

goal, however, is what we call the *literal* goal.[1] Literal goals are the goals of Gricean communication intentions, i.e., they are those goals that are intended to be achieved partly through recognition of the intention to achieve them. Thanks to Austin, Grice, Searle, and others, we have a fairly clear notion of what the literal goals of *illocutionary* acts are. For example, the literal goal of a promise is to let the hearer know that the speaker places himself under an obligation to do something. But it is not as obvious what the literal goal of a referring act is. Using the definition of referring that we have adopted, we take the literal goal of referring as making the hearer believe that it is mutually believed by all participants that both speaker and hearer have respective mental representations of the same object, and that a noun phrase is being used as a linguistic representation of that object.

Conditions of satisfaction. Illocutionary acts have *propositional content*, but referring acts do not. The propositional content of an illocutionary act determines what Searle [14] calls its *conditions of satisfaction*: a request that the door be opened is satisfied if and only if someone opens the door, and an assertion that the door is closed is satisfied (true) if and only if the door is indeed closed. But since a referring act lacks propositional content, it is not obvious what its conditions of satisfaction are.[2]

According to our theory of reference, the condition of satisfaction for a referring act is that the hearer have "identified" the referent. What counts as "proper identification" changes from discourse to discourse. For example, in "Replace *this 300-ohm resistor*," the hearer is asked to identify the referent in the sense of locating it in his visual field. But in "Tell me what other plays were written by *the author of Hamlet*," visual identification is clearly not required, although the hearer is still expected to identify the author of *Hamlet* in some other way. If the purpose of the referring act is 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 governing such concurrence. There is no simple way to state what such ground rules are, but they follow from the propositional content of the illocutionary act itself, general knowledge about the discourse situation and principles of rational behavior. For example, it follows from a theory of action that, if the speaker asks the hearer to pick up an object, the hearer can comply with the request only if he can perceive the object that he is about to manipulate. This fact implies at least one constraint on what the hearer must believe to have identified the referent: it must be something he can recognize perceptually. These ground rules, which change from one utterance to another, must be arrived at by analysis of what we call the *pragmatic* notion of referent identification.

[1] The term is taken from Kasher [9], by whom literal *purposes* are introduced. Our own use of the term, though, is slightly different.

[2] Note that specifying the conditions of satisfaction of a referring act is not the same as specifying its literal goal. The literal goal of a speech act and its conditions of satisfaction are usually distinct: If I tell you that I want the door closed and you understand me, the literal goal of my request is achieved. But it is still up to you whether or not to satisfy my request.

In summary, we have defined the literal goal and the conditions of satisfaction of referring in terms of what the hearer is supposed to *believe* and then *do*. The *literal goal* is to establish mutual belief that a particular object is being talked about. The referring act is *satisfied* when the hearer (1) knows the appropriate criteria for correct identification and (2) successfully satisfies these criteria.

3 A Formal Model of Referring

We now present a formal model of referring which satisfies the objectives set forth in Section 1.3. This requires that we (1) define the concept of an *individuating set*, which is essential to our theory, (2) specify a formal language in which axioms relevant to referring can be stated, (3) state the background assumptions of rationality upon which the theory rests, and (4) state axioms of referring.

3.1 Individuating Sets

We propose a structure called an *individuating set* as the formal entity upon which our theory of reference is based. Individuating sets are composed of intensional objects called *intensional object representations (IORs)*. A relationship of denotation can hold between each IOR and some individual in the world. An individuating set for an agent A is a maximal set of IORs, all believed by A to denote the same object. The set is *maximal* in the sense that if any two IORs are believed to be denoting the same object, they must belong to the same individuating set. The IORs themselves are either *perceptual IORs, discourse IORs,* or *functional IORs.* Perceptual IORs are mental representations of objects that result from perceptual acts (e.g., looking), and denote the object perceived. Discourse IORs are mental representations of objects that result from referring acts in discourse. Functional IORs are the values of functions whose arguments are also IORs. For example, a function *father-of*, may map IORs denoting persons onto IORs denoting their fathers.

It is important to note that an individuating set is the result of an agent's beliefs, not a mirror of what is actually the case. It is certainly possible to have an individuating set of terms that do not denote anything real (e.g., a child's representation of Santa Claus). Moreover, an agent may possess two distinct individuating sets for one and the same object (e.g., Oedipus's representations of his mother and his wife) or he may possess a single individuating set that, as a matter of fact, contains two IORs denoting different objects (e.g., when an agent fails to distinguish between Reagan and Regan). If all is well, and there actually is an object corresponding to all the terms in the set, we say that the *referent* of the latter is that object. If all is not well, the individuating set has no referent and the agent is simply mistaken or confused.

As we shall see, the individuating set makes it possible to state what one agent wants another to believe without stating the goal in terms of some common vocabulary of names. However,

individuating sets are motivated independently by a number of considerations, of which we can briefly mention only two. First, the concept of an individuating sets provides a solution to several problems that are raised by the referential/attributive distinction [12,11,10]. Second, Memory experiments [1,13] suggest that there are in fact structures in memory that encode the information contained in what we call an individuating set.

3.2 Identification constraints

As noted earlier, "identification", in the context of referring, is a *pragmatic* notion, not an epistemological one. Referent identification in conversation does *not* necessarily mean knowing who (or what) the referent is. In general, the requirements for referent identification can be characterized in terms of *constraints that apply to the relevant individuating set.* We call these constraints *identification constraints.*

If a hearer is cooperative, he forms an individuating set corresponding to each referring expression that is used by the speaker. We shall see later on how this is done. In the meantime, let IS be the hearer's individuating set that corresponds to the italicized referring expressions in the following examples. Here is how different identification constraints can be expressed as constraints on the set IS:

Example 1 Take *this chair* to my office.

In Example 1, the hearer should "identify" the chair in the sense of locating it in his visual field. We can express this requirement as a constraint on IS, namely, that it contain a perceptual IOR resulting from a perceptual action taking place in the current or some future situation.

Example 2 A friend of mine has just won $10,000 in a sweepstakes, but *the lucky bastard* will probably gamble it all away.

Identification constraint: that IS be "augmented" to include an IOR already introduced into the discourse. "Identification" here is simply making an anaphoric connection.

Example 3 *The man whose fingerprints these are*, whoever he is, must be insane.

Assume that the context of Example 3 is as follows. The speaker, investigating the gruesome murder of Smith, has just found clear fingerprints on what he believes to be the murder weapon. The speaker, of course, wishes to assert that whoever murdered poor Smith so brutally must be insane. Hence the identification constraint is that IS should contain the functional IOR corresponding to the description "Smith's murderer." This example illustrates why we insist that identification as the goal of the referring act is a pragmatic concept, rather than an epistemological one. Neither speaker nor hearer in this case has any idea who murdered Smith, so they cannot identify him in any epistemological sense. From a pragmatic point of view, however, there is a clear dichotomy:

if the hearer makes the connection between "the man whose fingerprints these are" and "Smith's murderer," he has identified the person the speaker is talking about. Otherwise he has not.

Example 4 I met *an old friend of mine* yesterday.

This is the case of the null set of identification constraints. IS contains a single discourse IOR meaning something like "whichever friend the speaker is talking about," and this is sufficient for pragmatic identification.

These examples show that the requirements for referent identification can be diverse indeed. Nevertheless, they can all be represented as constraints on the IORs comprising individuating sets.

3.3 A Formal Language

If one is to state the axioms of a theory of referring, then one needs a formal language in which to state them. We present such a formal language here. The logic we use is a dynamic modal logic of beliefs and goals whose semantics is very similar to Cohen's and Levesque's formal system [5]. We provide here a gloss of the various modal operators, predicates, and functions relevant to the theory of referring, and direct the reader desiring more information on the logic and its semantics to Cohen and Levesque's article.

Modal Operators:

- **Bel**(A, P): agent A believes that P.
- **MB**(A, B, P): A and B mutually believe P.
- **Goal**(A, P): agent A has goal P.
- **After**(E, P): after the occurrence of event E, P is true.

Predicates:

- Done(E): Event E has *just* occurred.
- Holds(d, x): d is a predicate of one argument, which is true of the individual x.

Terms and functions:

- Do(A, E): the event A's performing an action of type E.
- $e_1; e_2$: the event e_1, followed immediately by the event e_2.
- $p?$: an event that terminates if and only if p is true (e.g., Done$(p?; \text{Do}(Agent, Act))$ means that p was true immediately prior to $Agent$'s doing Act).

- t_u: A shorthand notation for the time at which the utterance u occurred. It is assumed that every situation has a discrete "time stamp."
- $cont(np)$: the descriptive content of np (a noun phrase).

Individuating sets:

- The function $\Delta(a, t, cont(d))$ maps agents, instances of time, and descriptive contents onto a *discourse* IOR. This IOR represents for agent a whatever individual the utterer of d had in mind when he used d at time t. Note that a may be either the speaker or the hearer.
- The function $\Pi(a, t, p)$ maps an agent, an instance of time, and a perceptual image onto a *perceptual* IORs.
- The function $IS(z)$ maps an IOR z onto the individuating set containing z.
- The function Ref(i) maps an individuating set i onto its referent. That is:

$$\text{Ref}(i) = \begin{cases} x & \text{if } \forall z \; z \in i \supset \text{denote}(z, x) \\ undefined & \text{otherwise} \end{cases}$$

- We shall use RefIS(z) as a simplified form of Ref(IS(z)).

3.4 Sincerity Helpfulness and Competence

There are a number of definitions and axioms, pertaining to a general theory of rational behavior, upon which rests both our theory of referring and Cohen and Levesque's theory of speech acts. Following Cohen and Levesque, we assume the following definitions of sincerity, helpfulness, and competence.

Sincerity: A is sincere to B with respect to P if whenever A has the goal of B believing P, he believes P himself. That is:

1 $\text{sincere}(A, B, P) \stackrel{def}{=} (\textbf{Goal}(A, \textbf{Bel}(B, P))) \supset \textbf{Bel}(A, P))$

Helpfulness: A is helpful to B with respect to P if whenever A believes B's goal is to get A to bring about P, A adopts the goal P.

2 $\text{helpful}(A, B, P) \stackrel{def}{=} (\textbf{Bel}(A, \textbf{Goal}(B, \textbf{Goal}(A, P) \supset \textbf{Goal}(A, P))$

Competence: A is competent with respect to P if, whenever A believes P, P is actually true.

3 $\text{competent}(A, P) \stackrel{def}{=} (\textbf{Bel}(A, P) \supset P)$

We also assume a number of properties about goals, belief, and mutual belief, which are stated in the appendix.

3.5 Referring Goals and Their Satisfaction

A theory of natural language must contain (among other things) a semantic component that assigns possible semantic interpretations to utterances, and a theory of speech acts and rationality that shows how uttering a sentence posessing certain features in the right circumstances is likely to affect the participants' mental states. We do not have much to say here about the semantic component, except that we assume it provides an analysis of utterances similar to Kamp's discourse representation theory [8]. Such an analysis hypothesizes discourse entitites, and provides constraints on how noun phrases are related to each other. In particular, it identifies certain noun phrases as potential referring expressions, which could be actual referring expressions if conditions on the mutual belief of the agents are satisfied in accordance with this theory of reference.

The following schema, which we call the *referring schema* is the cornerstone of our theory of referring. This schema states what is true of the hearer's beliefs about the speaker's and hearer's beliefs as a result of using *np* as a referring expression.

4 Referring Schema:

MB(S, H, sentence(U) $\wedge$ constituent-of(NP, U) $\wedge$ potential-refexp(NP)) $\supset$
After(Do(S, Utter(S, U)),
Bel(H, MB(S, H,
Goal(S, Bel(H,
$\exists z$ RefIS($\Delta(S, t_u, \text{cont}(NP))$) $= z \wedge$
Goal(S, Bel(H, Holds(cont(NP), RefIS($\Delta(S, t_u, \text{cont}(NP))$))))))))))

The referring schema states that if the speaker and hearer mutually believe that U is a sentential utterance and NP is both a constituent of U and a potential referring expression, then the speaker has the goal that the hearer believe two things: (1) that the speaker has an individuating set representing the individual corresponding to his utterance of NP, and (2) that the speaker wants the hearer to believe that the descriptive content of NP is true of the individual represented by his individuating set.[3]

Moreover, if indeed it is mutually believed that the speaker has a particular object in mind when he uses the noun phrase, it should also be mutually believed that the speaker intends the hearer to have a representation of that object. But in a typical Gricean fashion, once the hearer recognizes this intention, he *does* have such a representation. For example, if I recognize your intention that I should have a representation of whatever object you are talking about, then I *do* have such a representation, namely "Whatever object you are talking about." Thus we have what we call the Activation Axiom:

[3]Sometimes reference may be successful even though neither speaker nor hearer believe that the description is true of the referent (e.g. when irony is used). However, in such cases, the speaker is, in the sense implied by definition 1, insincere.

5 Activation Axiom:

$$\text{MB(S, H, } \exists x \text{ RefIS}(\Delta(S, t_u, \text{cont(np)})) = x) \supset$$
$$\text{MB(S, H, RefIS}(\Delta(S, t_u, \text{cont(np)})) = \text{RefIS}(\Delta(H, t_u, \text{cont(np)}))))$$

Note that, if the consequent of (5) is satisfied, the speaker has succeeded in activating in the hearer a mental representation of the object he (the speaker) has in mind, at least in part throngh the use of a noun phrase that is intended to be a linguistic representation of the same object. Since this is the absolute minimum required for a felicitous referring act, we can define the act of referring as the utterance of a noun phrase np by speaker S to hearer H at time t_u, with the goal being that the following mutual belief must hold:

6 $\text{MB(S, H, } \exists x \text{ [RefIS}(\Delta(S, t_u, \text{cont(np)})) = x \wedge \text{RefIS}(\Delta(H, t_u, \text{cont(np)})) = x]$

The goal of making (6) true is the *literal goal* of the referring act. For any referring action, (6) follows from the referring schema and the Activation Axiom *provided* that the hearer believes that it is mutually believed that the speaker is sincere, and competent with respect to the propostion that there is an individual represented by his individuating set.

Of course satisfaction of this literal goal is only the first step. Now the hearer is expected to find out what the identification constraints are and whether the newly created individuating set representing the referent satisfies those constraints. It is impossible to say anything in general about such constraints because they depend on the particular proposition that the speaker intends the hearer to believe. Identifying constraints derive from a variety of sources, including the syntactic and semantic properties of the utterance, the discourse, or general world knowledge [3]. Once recognized, however, they can be expressed in the logic as constraints on the relevant individuating set. For example, if visual identification is required, the identification constraint is that the individuating set contains a perceptual IOR, which can be expressed as follows:

7 $\exists t\,(t \geq t_u \wedge \exists z\ \Pi(H, t, z) \in \text{IS}(\Delta(H, t_u, \text{cont(np)})))$.

Statement (7) means that H's individuating set must contain a perceptual IOR resulting from a perceptual action occurring after the time of the utterance. If it is the case that (7) obtains, identification is complete, and the referring act is successful. If not, then the hearer has to devise an appropriate plan to bring about a situation in which (7) is true.

Of course, mere perception of the referent is not sufficient to satisfy the identification constraint. The hearer must believe that the thing he sees is the same as the object represented by the individuating set just introduced by the utterance. Although the exact process by which the hearer goes about making this determination is beyond the scope of this paper, it is safe to say that the descriptive content of the referring expression often plays an important role. If the hearer believes that it is mutually believed that the speaker is sincere and competent with respect to his goals,

and that he is sincere and competent with respect to the proposition that the referring description holds of the referent, then it follows from the referring schema that the speaker and hearer mutually believe that the description is true of the referent. This knowledge in turn is useful for identifying the referent.

If the hearer does not believe the speaker is competent with respect to the referring description (perhaps he believes that nothing satisfies the description), the identification constraints can still be satisfied. Goodman [7] discusses in detail a process by which identification can proceed in such circumstances.

The appendix illustrates precisely how the recognition of the literal goal of a referring act and the identification conditions follows from the referring schema, activation axiom, and Cohen and Levesque's schema for the effects of the utterance of an imperative sentence.

According to this analysis of referring, the satisfaction of the literal goal of referring does not depend on the descriptive content of the referring expression. The descriptive content plays a role only in the satisfaction of identification constraints, and because this can be carried out even if the hearer does not believe the description, reference can succeed even if the speaker uses an "incorrect" description.[4]

3.6 Referring as Informing and Requesting

Suppose that a speaker utters a referring expression with descriptive content d, and that the hearer is able to satisfy any relevant identification constraints independent of knowing the individuals of which d may be true. If the hearer still assumes the speaker's sincerity, and his competence with respect to the propostion that d holds of the referent, then it follows from the Referring Schema that the hearer believes that d holds of the referent. Thus, the conclusion that follows from (4) is

8 **Bel**(H, **MB**(S, H,
Goal(S, **Bel**(H,
Goal(S, **Bel**(H, Holds(cont(NP), RefIS($\Delta(S, t_u, \text{cont(NP)})$))))))))),

which is precisely the belief that characterizes an informing action. An agent who is planning utterances can exploit the independent satisfaction of identification constraints to inform the hearer of properties of the referent through an appropriate choice of referring expression. In situations in which identification constraints are minimal, one would predict that referring acts can subsume informing acts very freely, and this does indeed seem to be the case, as evidenced by the equivalence between "I met someone yesterday. He was an old friend from high school." and "I met an old friend from high school yesterday."

[4] The descriptive content surely plays a role in recognizing that the noun phrase is a potential referring expression. But as we have seen, once the noun phrase is taken as a potential referring expression, the descriptive content plays no role in establishing that both speaker and hearer have mental representations of the same object.

This analysis of referring expressions also predicts situations in which a referring expression can be taken as a request to perform an action associated with the satisfaction of the identification constraints. As noted earlier (p. 4) Cohen argues that referring actions can be regarded as requests that the hearer identify the referent. According to our analysis, reference in general does not entail a request, because in situations in which no identification constraints apply, or in which the satisfaction of identification constraints follows from the speaker and hearer's shared knowledge, the hearer doesn't have to do anything for the referring goal to be satisfied. However, in those situations in which it is mutually believed that the satisfaction of the identification constraints necessarily entails the performance of a particular action it follows from our analysis that the utterance is in fact a request to perform the identifying action. Suppose the speaker requests the hearer to replace a certain object, and suppose that the conditions of sincerety, and compentence with respect to the existence of the referent are satisfied. Then the Referring Schema and the effects of the imperative utterance entail the following proposition:

9 **Bel**(H, **MB**(S, H,
Goal(S, **Bel**(H,
Goal(S, **Goal**(H, Done(Do(H, replace($\Delta(S, t_u, cont(NP)$)))))))))

Now, if speaker and hearer mutually believe that in order for the hearer to get an object replaced, he must look in a certain location L, then in any situation in which replacement has been performed, it must be the case that an action of looking had been done. Therefore, the following belief holds:

10 **Bel**(H, **MB**(S, H,
Goal(S, **Bel**(H,
Goal(S, **Goal**(H, Done(Do(H, Look(L)))))))))

Note that 10 is precisely the belief that characterizes a request to perform the action required to identify the referent visually. In the task oriented domains examined by Cohen, these assumptions obtain for the vast majority of the referring actions.

4 Conclusion

In this paper we have presented a formal model of referring. It has been shown that referring, like other speech acts, has a *literal goal* and *criteria for success*. Satisfaction of the literal goal establishes mutual belief concerning the speaker's intention to refer to a particular object. The referring act succeeds when the object is identified correctly.

An important feature of this analysis is that most of its parts are independently motivated. The theory of speech acts and rationality was developed by Cohen and Levesque for reasons other than

accounting for referring actions; individuating sets are motivated independenly of this analysis, and the referring schema can be justified independently of its role in showing how referring expressions are used for requesting and informing.

Our model of referring itself depends essentially on two concepts: *individuating sets,* and *identification constraints.* These concepts allow us to describe successful reference without requiring that all objects have standard names or that agents should know these names. Using Cohen and Levesque's dynamic modal logic, we have formalized the referring goals and stated axioms that relate individuating sets to utterances as well as to other actions performed by agents. Such axioms could be utilized by language-processing systems for either recognizing or planning speech acts that involve referring expressions. Moreover, the incorporation a theory of referring into a general theory of speech acts and rationality makes it possible to understand how referring and other speech acts are related. In particular, we can explain how referring acts can be used to achieve multiple goals — i.e., to refer and to provide information about a referent, or to refer and to request that an identifying action be performed.

Acknowledgements

This research was supported by the National Science Foundation under grant DCR-8407238. The authors are grateful to Phil Cohen, Asa Kasher, Ray Perrault, and Martha Pollack for comments on earlier drafts of this paper.

References

[1] John R. Anderson. The processing of referring expressions within a semantic network. In *Proceedings of TINLAP-2*, pages 51–56, 1978.

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

[3] Douglas E. Appelt. Reference and pragmatic identification. In *Theoretical Issues in Natural Language Processing-3*, New Mexico State University Memoranda in Computer and Cognitive Science, 1987.

[4] Philip R. Cohen. Referring as requesting. In *Proceedings of the Tenth International Conference on Computational Linguistics*, pages 207–211, 1984.

[5] 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.

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

[7] Brad Goodman. Repairing miscommunication: relaxation in reference. In *Proceedings of the National Conference on Artificial Intelligence*, pages 134–138, American Association for Artificial Intelligence, 1983.

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

[9] Asa Kasher. What is a theory of use? *Journal of Pragmatics*, 1:105–120, 1977.

[10] 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.

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

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

[13] Andrew Ortony and R. Anderson. Definite descriptions and semantic memory. *Cognitive Science*, 1:74–83, 1977.

[14] John Searle. *Intentionality: An Essay in the Philosophy of Mind.* Cambridge University Press, Cambridge, England, 1983.

[15] John Searle. Referential and attributive. In *Expression and Meaning: Studies in the Theory of Speech Acts*, Cambridge University Press, Cambridge, England, 1979.

5 Appendix

This appendix illustrates how the literal goal of referring and the identification constraints follow from the axioms of reference and basic assumptions of the theory of speech acts and rationality. Suppose S tells H under appropriate circmstances: "Please replace the 300-ohm resistor," and let $\lambda x.300\Omega R(x)$ be the descriptive content of the noun phrase "the 300 ohm resistor." We show how S's referring goals are satisfied by showing how H's beliefs and goals change after S's utterance.

Elementary axioms. The following elementary properties of goals, beliefs and mutual belief are necessary for the derivation, and are assumed to be mutually believed by the speaker and hearer.

11 $\mathbf{Bel}(A, \mathbf{Bel}(A, P)) \supset \mathbf{Bel}(A, P)$

12 $\mathbf{Bel}(A, P \wedge Q) \supset (\mathbf{Bel}(A, P) \wedge \mathbf{Bel}(A, Q))$

13 $(\mathbf{Bel}(A, P \supset Q) \wedge \mathbf{Bel}(A, P)) \supset \mathbf{Bel}(A, Q)$

14 $\mathbf{Goal}(A, P \wedge Q) \supset (\mathbf{Goal}(A, P) \wedge \mathbf{Goal}(A, Q))$

15 $\mathbf{Bel}(A, \mathbf{Goal}(A, P)) \supset \mathbf{Goal}(A, P)$

16 $(\mathbf{Goal}(A, P(x)) \wedge \mathbf{Bel}(A, x = y)) \supset \mathbf{Goal}(A, P(y))$

17 $(\mathbf{MB}(S, H, P \supset Q) \wedge \mathbf{MB}(S, H, P)) \supset \mathbf{MB}(S, H, Q)$

18 $\mathbf{After}(E, P) \supset (\mathrm{Done}(E) \supset P)$

Because the propositional content of the request is needed to derive identification conditions, we combine the Referring Schema with the axiom describing the effects of the utterance of an imperative sentence given by Cohen and Levesque.

19 *Referring within a request:*
MB*(S, H, (all preparatory conditions are satisfied))* ⊃
After*(*Do*(S,* utter*(S, "Please replace the 300-ohm resistor")),*
Bel*(H,* **MB***(S, H,* **Goal***(S,* **Bel***(H,*
$\exists z$ RefIS$(\Delta(S, t_u, \lambda x.300\Omega R(x))) = z) \wedge$
Goal*(S,*
Bel*(H,* Holds$(\lambda x.300\Omega R(x),$ RefIS$(\Delta(S, t_u, \lambda x.300\Omega R(x)))) \wedge$
Goal*(H,* Done*(*Do*(H,* replace(RefIS$(\Delta(S, t_u, \lambda x.300\Omega R(x))))))))))))$

20 done(utter(S, "Please replace the 300-ohm resistor") Assumption

Assuming that it is mutually believed that all preparatory conditions are satisfied we get:

21 **Bel**(H, **MB**(S, H, **Goal**(S, **Bel**(H,
$\exists z$ RefIS$(\Delta(S, t_u, \lambda x.300\Omega R(x))) = z) \wedge$
Goal(S,
Bel(H, Holds$(\lambda x.300\Omega R(x),$ RefIS$(\Delta(S, t_u, \lambda x.300\Omega R(x)))) \wedge$
Goal(H, Done(Do$(H,$ replace(RefIS$(\Delta(S, t_u, \lambda x.300\Omega R(x))))))))))$ 20, 18, and 19

22 **Bel**(H,**MB**(S, H, **Bel**(S,
$(\exists z$ RefIS$(\Delta(S, t_u, \lambda x.300\Omega R(x))) = z) \wedge$
Goal(S,
Bel(H, Holds$(\lambda x.300\Omega R(x),$ RefIS$(\Delta(S, t_u, \lambda x.300\Omega R(x)))) \wedge$
Goal(H, Done(Do$(H,$ replace(RefIS$(\Delta(S, t_u, \lambda x.300\Omega R(x))))))))))$
21, 1(sincerity), 17, and 13

23 **Bel**(H, **MB**(S, H,
Bel(S, $\exists z$ RefIS$(\Delta(S, t_u, \lambda x.300\Omega R(x))) = z))))$ 22,12,17, and 13

24 **Bel**(H, **MB**(S, H,
Bel(S, **Goal**(S,
Bel(H, Holds($\lambda x.300\Omega R(x)$, RefIS($\Delta(S, t_u, \lambda x.300\Omega R(x))$)) $\wedge$
Goal(H, Done(Do(H, replace(RefIS($\Delta(S, t_u, \lambda x.300\Omega R(x))$)))))))))
22,12,17, and 13.

25 **Bel**(H, **MB**(S, H, **Goal**(S,
Bel(H, Holds($\lambda x.300\Omega R(x)$, RefIS($\Delta(S, t_u, \lambda x.300\Omega R(x))$)) $\wedge$
Goal(H, Done(Do(H, replace(RefIS($\Delta(S, t_u, \lambda x.300\Omega R(x))$)))))))))
24, 15, 17, and 13.

26 **Bel**(H, **MB**(S, H, **Goal**(S, **Bel**(H,
Holds($\lambda x.300\Omega R(x)$, RefIS($\Delta(S, t_u, \lambda x.300\Omega R(x))$)))))))
25, 14, 17, and 13.

27 **Bel**(H, **MB**(S, H, **Goal**(S, **Goal**(H,
Done(Do(H, replace(RefIS($\Delta(S, t_u, \lambda x.300\Omega R(x))$)))))))))
25, 14, 17, and 13.

28 **Bel**(H, **MB**(S, H, $\exists z$ RefIS($\Delta(S, t_u, \lambda x.300\Omega R(x))) = z$))
23, 3(S's competence), 17, and 13.

29 **Bel**(H, **MB**(S, H, RefIS($\Delta(S, t_u, \lambda x.300\Omega R(x))$) = RefIS($\Delta(H, t_u, \lambda x.300\Omega R(x))$)))
28, 5(Activation Axiom), 17, 13.

Note that at this point, the literal goal of the referring act has been satisfied: if the hearer is right, it is mutually believed that both speaker and hearer have a mental representation of the very same object.

30 **Bel**(H, **MB**(S, H, **Goal**(H,
Done(Do(H, replace(RefIS($\Delta(S, t_u, \lambda x.300\Omega R(x))$)))))))
27, 2(helpfulness), 17, and 13.

31 **Bel**(H,**MB**(S, H, **Goal**(H,
Done(Do(H, replace(RefIS($\Delta(H, t_u, \lambda x.300\Omega R(x))$)))))))
30, 29, 16(substitution), 17, and 13.

Thus, H can assume that he is expected to replace the object determined by the newly created individuating set. However, he still does not know which object it is. What is being presupposed, at this point, is a theory of physical actions. Such a theory is likely to state that an executable procedure for replacing anything requires that the agent *perceives* the object of the action. This means that the individuating set must contain a perceptual term:

32 $\forall a, \phi(\mathrm{Do}(A, \mathrm{replace}(\mathrm{Ref}(\phi)))) \equiv (\exists z \exists t\ \mathbf{Bel}(A, \Pi(A, t, z) \in \phi)?;$
$\mathrm{Do}(A, \mathrm{change}(\mathrm{Ref}(\phi)))$

Applying 32 to 31 we get 33:

33 **Bel**(H, **MB**(S, H, **Goal**(H,
Goal(H, Done(
$(?\exists z \exists t\ \mathbf{Bel}(H, \Pi(H, t, z) \in \mathrm{RefIS}(\Delta(H, t_u, \lambda x.300\Omega R(x)))));$
$\mathrm{Do}(H, \mathrm{change}(\mathrm{RefIS}(\Delta(H, t_u, \lambda x.300\Omega R(x))))))$ 32, 31

Thus, the identification constraint on the individuating set is that it contain a perceptual term. Since at this stage it does not, the hearer has to get one, i.e., perform a perceptual action. It is tempting to say at this point that the hearer should simply use the description "The 300-ohm resistor" to locate the resistor and aquire the right perceptual term. However, this would be a gross over simplification. Identification involves a host of assumptions about how accurate the description is, how inaccuracies are corrected [7], how mutual belief about the properties of objects in the scene is achieved, and so on. The actual satisfaction of identification constraints is an extremely important area of research, and understanding it is crucial to both the planning and the understanding of referring expressions. A discussion of these matters, however, is beyond the scope of this paper.