A Commitment-Based Approach to Agent Speech Acts and Conversations

Marco Colombetti

Department of Electronics and Information
Politecnico di Milano
Piazza Leonardo da Vinci 32, I-20133 Milano
Tel. (+39) 02 2399 3686 – Fax (+39) 02 2399 3411
marco.colombetti@polimi.it

Faculty of Communication Sciences
Università della Svizzera italiana
Via Ospedale 13, CH-6900 Lugano
Tel (+41) 91 912 4767 – Fax (+41) 91 912 4647
marco.colombetti@lu.unisi.ch

ABSTRACT

This paper presents the main elements of Albatross, an agent communication language whose definition is currently under development. The semantics of Albatross, based on the social notion of commitment, allows one to define speech act types in a neat and concise way. I describe the logical relationship between messages and speech acts; give sample definitions of declarative, assertive, commissive, and directive speech acts; discuss the relationship between speech acts, normative systems, and rational action; and suggest a way of dealing with agent conversations.

Keywords

Agent communication languages, speech acts, commitment, rationality, normative systems, agent conversations.

1. INTRODUCTION

In recent years, the importance of defining a standard framework for agent communication has been widely recognized. This scientific endeavor has several facets: agent communication requires at least the definition of a standard agent communication language (ACL) and of common conversational protocols. So far, at least one ACL has been extensively used in experimentation (KQML [8]), and a second language has been proposed as a possible standard (FIPA ACL [9]). Both proposals are based on a notion of speech act close to the concept of illocutionary act as developed in speech act theory [3,14,16].

Many researchers are not satisfied with existing ACLs. A first difficulty is that there is no general consensus on what a formal semantics for such languages should look like. Some attempts to define semantics for ACLs are based on mental states [6,9,11]; some are based on the social notion of commitment [17,18]; and some propose to equate the meaning of a speech act with the set of allowable responses [13]. Another problem is that the relationship between speech acts and various related entities (like agent mental states, conversations, and so on) is not completely understood yet. For example, the protocols of FIPA ACL form a sort of dialogue grammar that regulates the speech acts that can be performed by an agent during a conversation; however, the relationship between the semantics of a speech act and the role it can play in a protocol is not completely clarified.

In this paper I shall sketch the definition of a new speech-act based ACL, which I named *Albatross* (*A*gent *l*anguage *ba*sed on a *tr*eatment *o*f *s*ocial *s*emantics). The language has a simple semantics, based on the social notion of commitment, which appears to have remarkable advantages over the more traditional proposals based on mental states (see for example [5,6,11]). In this respect, my current proposal also departs from previous work by myself and colleagues [1], although other aspects of Albatross rely substantially on such work. The reasons why I prefer an approach based on commitments are diverse, and have been already discussed in the literature (see for example [17,18]). The main advantage is that commitments, contrary to mental states, are objective and public. They do not need to be reconstructed and attributed to other agents through inference processes, and can be recorded in a public store for further reference. A second advantage is that commitments provide for a principled way to connect speech acts to the internal world of individual rationality and to the external world of conversational protocols. Moreover, they allow for a natural treatment of the legal consequences of speech acts, which is especially important for certain kinds of agents, like those involved in electronic commerce.

As we shall see, my treatment is based on commitments of two different sorts. Given that agents are assumed to be autonomous, at least to a certain degree, they have the power to commit to what they like, but in general they cannot directly commit another agent to anything. However, an agent can propose that another agent make a given commitment. In the terminology used in this paper, this amounts to putting another agent in a state of *precommitment*. The use of precommitments allows for a simple treatment of directive speech acts and conversational rules.

This paper is structured as follows. In Section 2, I introduce the basic elements of the logical language used to define the semantics of Albatross. In Sections 3, I show how different kinds of Albatross speech acts can be defined. In Section 4, I discuss some relationships between the social semantics of speech acts, individual rationality, and normative systems. In Section 5, I suggest a way of dealing with agent conversations

through conversational commitments. Finally, in Section 6 I draw some conclusions and outline directions for future work.

2. THE LOGICAL LANGUAGE

The definition of Albatross is based on an extended first order modal language **L**, which I simply call *the logical language*. Given that this research is currently in progress, the formal definition of **L** is admittedly incomplete.

2.1 Terms and denotations

L is a first order modal language with terms of different sorts, including at least: agent, action token, action type, date, force indicator, and message body. A model M for **L** is built from a nonempty set W of possible worlds, a nonempty domain D_s for every sort s, and additional functions and relations that will be defined in the sequel. At each world, every term of sort s is assigned a denotation by the function δ: $\mathbf{T}_s \times W \longrightarrow D_s$, where $\mathbf{T}_s$ is the set of all terms of sort s, and D_s is the domain of individuals of sort s. As usual, the denotation function is defined recursively from the interpretation of constants and functors, and from the evaluation of variables. For terms that are rigid designators, the denotation function is constant with respect to its second argument.

2.2 Time

Possible worlds are thought of as complete snapshots ordered by discrete time instants. The function *next*: $W \longrightarrow 2^W$ associates to every possible world the set of possible worlds that come next in time. A *w-path* is an infinite sequence $(w_n)_{n\in\mathbf{N}}$ of worlds such that $w_0 = w$ and $w_{n+1}\in next(w_n)$ for every n. Every possible world is assigned an absolute time (i.e., an integer number) by a function *time*: $W \longrightarrow \mathbf{Z}$. If $w'\in next(w)$, then $time(w') = time(w)+1$. Terms of sort date are interpreted onto absolute times. The temporal operator *Next* has the following semantics:

$M,w \models Next\ \varphi$ iff
for every w', if $w' \in next(w)$, then $M,w' \models \varphi$.

In this paper I shall use a further temporal operator, *Within*. If d is a term of sort date, a statement of the form $Within(d,\varphi)$ means that for every future development of the current world, φ holds at some world not later than date d:

$M,w \models Within(d,\varphi)$ iff
all *w-paths* contain a world w'
such that $time(w') \le \delta(d,w)$ and $M,w' \models \varphi$.

Finally, I shall use the constant *now*, of sort date, with the following interpretation:

$\delta(now,w) = time(w)$.

2.3 Actions

If e is a term of sort action token, and α is a term of sort action type, then $Act(e,\alpha)$ means that e is a token of action type α. For example, suppose that the term $buy(x,y,i,p)$ represents the action type "agent x buys item i at price p from agent y." Then, $Act(e,buy(x,y,i,p))$ means that e is a token of such a type.

If e is a term of sort action token, and x is a term of sort agent, then $Done(e,x)$ means that e has just been completed by x. In such a case, I say that x is the *actor* of e. An axiom states that every action token has at most one actor:

$Done(e,x) \wedge Done(e,y) \rightarrow x = y$.

To make logical statements more concise, I shall "overload" predicate *Done* according to the following definitions:

$Done(e,x,\alpha) =_{def} Act(e,\alpha) \wedge Done(e,x)$,
$Done(x,\alpha) =_{def} \exists e\ Done(e,x,\alpha)$,
$Done(x,\alpha/d) =_{def} Within(d,Done(x,\alpha))$.

Action types can be combined through various operators to give new action types. In this paper I shall only use the disjunction operator, '|'. If α and β are two action types, the expression $(\alpha \mid \beta)$ denotes the disjunction of α and β. Its semantics is defined by the following axiom:

$Act(e,(\alpha \mid \beta)) \leftrightarrow Act(e,\alpha) \vee Act(e,\beta)$.

2.4 Messages

The concrete syntax of Albatross messages will not be specified. I shall only assume that a message is an expression with subexpressions specifying a *sender*, a list of *receivers*, a *force indicator* (in the sense of speech act theory), and a *body* (i.e., a statement of a *content language* conveying the content of the message). If x and y are terms of sort agent, f is a term of sort force indicator, and s is a term of sort message body, the term $send(x,y,f,s)$ denotes the following action type: a message is sent with sender x, y as one of the receivers, force indicator f, and body s. I require that only the agent identified as the sender of a message can possibly send the message:

$Act(e,send(x,y,f,s)) \wedge Done(e,z) \rightarrow x = z$.

I also take it that the semantics of Albatross messages can be expressed in **L**. More precisely, I assume that for every message body s there is a logical statement φ such that $Holds(s) \leftrightarrow \varphi$ is valid, where the intuitive meaning of $Holds(s)$ is that s holds. This assumption cannot be expressed in a first order language, and is to be understood as metatheoretic. Note that I do not suppose that every formula of **L** has a corresponding statement in the content language: the logical language can, and presumably will, be more expressive than the content language.

2.5 Speech acts

A speech act (in the restrictive sense of an illocutionary act) has four components: a *speaker*, a set of *addressees*, a *force*, and a *content*. A term of the form $speechAct(x,y,f,\varphi)$ denotes the following action type: a speech act is performed with x as the speaker, y as one of the addressees, force f, and content φ. The speaker of a speech act coincides with the agent that performs it:

$Act(e,speechAct(x,y,f,\varphi)) \wedge Done(e,z) \rightarrow x = z$.

An unusual feature of the term $speechAct(x,y,f,\varphi)$ is that one of its arguments is not itself a term, but a logical statement. This reflects the fact that speech acts, unlike other types of action,

have a *propositional content*. This feature is likely to make the logical language difficult to manage, unless the formulae that appear as contents of speech acts are suitably constrained. Given that such contents derive from message bodies, constraints may be implemented by limiting the expressive power of the content language. In any case, the semantics of functional terms containing statements as arguments can be defined as follows. In general, the denotation of a functional term $f(t_1,...,t_n)$ is obtained by applying $[f]$, the interpretation of f, to the denotations of t_1, ..., t_n. As the denotation of a logical statement, φ, we can take its *truth set* in M, $\|\varphi\|_M$, defined as:

$$\|\varphi\|_M = \{w: M,w \models \varphi\}.$$

Now assume that f is a function symbol with n+1 arguments, such that: (i), the first n arguments are terms denoting individuals in the domains D_1, ..., D_n; (ii), the n+1-*th* argument is a statement; and (iii), the term $f(t_1,...,t_n,\varphi)$ denotes an individual in domain D. We can interpret f onto a function

$$[f]: D_1 \times ... \times D_n \times 2^W \longrightarrow D.$$

We then define the denotation of $f(t_1,...,t_n,\varphi)$ in M at world w as follows:

$$\delta(f(t_1,...,t_n,\varphi),w) = [f](\delta(t_1,w),...,\delta(t_n,w),\|\varphi\|_M).$$

With this semantics, the following inference rule of extensionality is valid:

$$(\mathrm{RE}_f)\ \frac{\varphi \leftrightarrow \psi}{f(t_1,...,t_n,\varphi) = f(t_1,...,t_n,\psi)}.$$

We are now ready to define the relationship between messages and speech acts. I want to express the fact that sending a message with certain features counts as a speech act with features related to the features of the message. As the link between a message body and its logical representation is given by a valid formula, the relationship between messages and speech acts has to be expressed through an inference rule:

$$(\mathrm{RS})\ \frac{Holds(s) \leftrightarrow \varphi}{Act(e,send(x,y,f,s)) \rightarrow Act(e,speechAct(x,y,f,\varphi))}.$$

2.6 Commitment

Commitment is a deontic notion, akin to obligation. Therefore, to formalize commitment we can take inspiration from deontic logic (see for example [2]). It seems to me that, with respect to a broad notion of obligation, the distinctive features of commitment are the following:

- Any commitment is always a commitment *to* some state of affairs or course of action, *taken by* some agent (which, following [19], I shall call the *debtor* of the commitment) *relative to* some other agent or set of agents (the *creditors* of the commitment).
- A commitment arises as the effect of performing some action.
- A commitment persists in time until it is cancelled. A commitment can be cancelled in a variety of ways, in particular by being fulfilled, rescinded, or otherwise modified through renegotiation or delegation. This aspect is not analyzed in the current paper.

A statement of the form $C(e,x,y,\varphi)$ means that action e commits agent x to φ relative to agent y. The symbol C is similar to a family of modal operators, indexed by the first three arguments. However, given that such arguments are arbitrary terms of the logical language (of suitable sorts), I prefer to call it a *modal predicate*.

The C predicate expresses a deontic concept, and is going to play a role analogous to that of the obligation operator in deontic logic. Unfortunately, choosing a suitable logic for a deontic operator is not an easy task. Classical modal logic is plagued with technical difficulties, mainly related to the treatment of conditional and conflicting obligations. My approach here is to choose the weakest possible logic for C, that is, the logic induced by minimal models (see [4]). The semantics of commitment will therefore be defined though a function f_C: $D_{\mathrm{action}} \times D_{\mathrm{agent}} \times D_{\mathrm{agent}} \times W \longrightarrow 2^{2^W}$, by stipulating that:

$$M,w \models C(e,x,y,\varphi) \quad \text{iff} \quad \|\varphi\|_M \in f_C(\delta(e,w),\delta(x,w),\delta(y,w),w).$$

This semantics enforces the validity of the rule of extensionality:

$$(\mathrm{RE}_C)\ \frac{\varphi \leftrightarrow \psi}{C(e,x,y,\varphi) \leftrightarrow C(e,x,y,\psi)}.$$

The resulting logic of C is indeed very weak. Nevertheless, I shall show in Section 4 that one can reason on commitments in an interesting way by using the concept of violation.

As already pointed out, I also need a weaker version of commitment, which I call *precommitment*. The statement $PC(e,x,y,\varphi)$ means that e precommits agent x to φ relative to agent y. The semantics of $PC(e,x,y,\varphi)$ is defined analogously to the semantics of C.

Another important feature of commitments is that they persist in time until they are cancelled. However, I have not yet developed an exhaustive model of the ways in which a commitment can be cancelled. In the rest of this paper, therefore, commitments and precommitments should be intuitively understood as persisting in time until they are explicitly negated.

3. SPEECH ACTS

In this section I define a number of speech acts. I follow John Searle [15] in classifying illocutionary acts into five categories: declarations, assertives, commissives, directives, and expressives. All these categories are relevant to agent communication, with the exception of expressives: at least in the foreseeable future, agents are not likely to spend much time in congratulating, apologizing, and so on. For each of the remaining four categories, I shall first define a basic act, that is, a sort of zero-point for the category; then I shall introduce more complex speech acts to show the flexibility of commitment-based semantics.

3.1 Declarations

A declaration is a speech act that brings about a state of affairs that makes its content true. Examples of declarations are "the auction is open" (used to open an auction) and "the price of this item is 100 euros" (used to set the price of the item). It is not difficult to see that various types of agents (e.g., those involved in electronic commerce) should be able to perform declarations.

The point of a declaration is to bring about the truth of what is declared. However, in order for a declaration to be effective, the declaring agent must be endowed with specific powers. For example, an auction cannot be open by one of the participants, nor can the price of an item be set by a client. When agent x is empowered to bring about φ by declaration, I write $Empowered(x,\varphi)$. Here are the axioms for declarations:

(Declare1) $declare(x,\varphi) = speechAct(x,y,declare,\varphi)$,

(Declare2) $Done(x,declare(x,\varphi)) \wedge Empowered(x,\varphi) \rightarrow \varphi$.

Most of my definitions of speech acts will have this structure, so let me briefly analyze it. The first axiom, Declare1, defines a new action type, represented by terms of the form $declare(x,\varphi)$. This action type coincides with the subclass of all speech acts whose force is '*declare*.' Together with rule RS, axiom Declare1 tells us that sending a message with '*declare*' as the force indicator is a means to realize a declaration. The second axiom, Declare2, specifies the effect of performing a declaration in a context in which the condition $Empowered(x,\varphi)$ holds. In principle, it would be simpler to specify the effect of a declaration without defining a new action type, for example by adopting the axiom:

$Done(x,speechAct(x,y,declare,\varphi)) \wedge Empowered(x,\varphi) \rightarrow \varphi$.

However, introducing a specific notation for each speech act type is useful for the definition of new speech acts (see for example Sections 3.7 and 3.8).

Axiom Declare2 states that if an empowered agent performs a declaration, the declared state of affairs is brought about. Of course, there are limitations to the declaration powers that can be attributed to agents. An agent may well open an auction or set the price of a good, but it cannot, for example, fix the CPU of your computer by simply declaring that the CPU has been fixed. However, a formal analysis of empowerment is beyond the scope of this paper (see [10] for an interesting proposal).

3.2 Assertives

The point of an assertive act is to commit its actor to the truth of what is asserted, relative to every addressee. The basic assertive act, that is, the act of *asserting*, can be defined as follows:

(Assert1) $assert(x,y,\varphi) = speechAct(x,y,assert,\varphi)$,

(Assert2) $Done(e,x,assert(x,y,\varphi)) \rightarrow C(e,x,y,\varphi)$.

Note that, contrary for example to FIPA ACL, I do not take *informing* as the basic assertive act. In contrast with asserting, an act of informing presupposes that the speaker believes its content and intends the addressees to believe it as well. Conditions of this type are not considered in the commitment-based semantics of Albatross. FIPA ACL also defines additional assertives, like *confirming* and *disconfirming*. It is still not clear to me whether such acts are going to play an important role in agent communication. In any case, in a commitment-based language they can be dealt with as operations on commitments. For example, under certain conditions an agent may be empowered to retract a previous commitment and to replace it with a new one. This operation can be regarded as the commitment-based counterpart of disconfirming. However, a much more detailed analysis of commitment is necessary before we can give a complete treatment of such operations.

3.3 Commissives

The point of a commissive act is to commit the speaker, relative to every addressee, to the execution of an action of a given type within a limiting date. Here is the definition of the basic commissive act, *promising*:

(Promise1) $promise(x,y,\varphi) = speechAct(x,y,promise,\varphi)$,

(Promise2) $Done(e,x,promise(x,y,Done(x,\alpha/d))) \wedge d > now \rightarrow C(e,x,y,Done(x,\alpha/d))$.

It appears from this definition that a commissive can be replaced by an assertive: instead of promising to do α, an agent might assert that it will do α. It seems to me that this is a consequence of "projecting" the complex concept of a speech act onto the single dimension of commitment. In fact, in Searle's speech act theory assertives and commissives differ with respect to conditions that are not taken into account in the semantics of Albatross. In any case, I prefer to maintain the distinction between assertives and commissives, because of its intuitive appeal.

3.4 Directives

The point of a directive act is to have the addressee or addressees perform some action within a limiting date. In a mentalistic approach to communication, one typically treats directives in terms of intentions. Coherently with the social standpoint advocated in this paper, I deal with directives in terms of commitments. However, as already pointed out, an agent cannot directly bring about a commitment of another agent. To solve this problem, I rely on the concept of precommitment.

The basic directive act, *requesting*, can be defined as follows:

(Request1) $request(x,y,\varphi) = speechAct(x,y,request,\varphi)$,

(Request2) $Done(e,x,request(x,y,Done(x,\alpha/d))) \wedge d > now \rightarrow PC(e,y,x,Done(x,\alpha/d))$.

3.5 Accepting and rejecting

In general, a precommitment can be accepted or rejected by its debtor. *Accepting* is a speech act that transforms a precommitment, typically deriving from a directive act, into a full commitment. It can be defined as follows:

(Accept1) $accept(x,y,\varphi) = speechAct(x,y,accept,\varphi)$,

(Accept2) $Done(x,accept(x,y,\varphi)) \wedge PC(e,x,y,\varphi) \rightarrow C(e,x,y,\varphi)$.

On the contrary, an act of *rejecting* cancels a precommitment:

(Reject1) $reject(x,y,\varphi) = speechAct(x,y,reject,\varphi)$,

(Reject2) $Done(x,reject(x,y,\varphi)) \wedge PC(e,x,y,\varphi) \rightarrow Next\ \neg PC(e,x,y,\varphi)$.

3.6 Contracts

Intuitively, we feel that at least in some cases a directive can actually generate a commitment of an addressee. The most evident case is when the directive has the force of an order. However, the nondefective performance of similar directives presupposes the existence of a suitable social relationship. In the case of orders, for example, a relationship of subordination must be established between the speaker and the addressees. It seems to me that similar cases might well occur with artificial agents. For instance, we can think of a situation in which an agent is bound to accept a precommitment because of a previous agreement in this sense. This kind of pre-existing agreements I shall call *contracts*. More precisely, I assume that agents x and y may be bound by a contract to accept all precommitments of x, relative to y, to do an action of type α, and I express this through the statement $Contract(x,y,\alpha)$. We have the following axiom:

(Contract) $Contract(x,y,\alpha) \wedge PC(e,x,y,Done(x,\alpha/d)) \wedge d > now \rightarrow C(e,x,y,Done(x,\alpha/d))$.

To show the flexibility of commitment-based semantics, I shall now define a few more types of speech acts, namely yes-no questions, wh-questions, and proposals.

3.7 Yes/no questions

Yes/no questions are a notable subclass of directive acts, by which an agent requests another agent to assert whether some sentence is true or false. I assume that answers to yes/no questions must be given within k instants, where k may be viewed as a global constant, or as a further parameter of questions:

(AssertIf) $assertIf(x,y,\varphi) = (assert(x,y,\varphi) \mid assert(x,y,\neg\varphi))$,

(AskIf) $askIf(x,y,\varphi) = request(x,y,Done(y,assertIf(y,x,\varphi)/now+k))$.

3.8 Wh-questions

Before defining wh-questions, it is necessary to establish what counts as a valid wh-answer. A *wh-answer* is an assertive act whose content is a *designating statement*, that is, a statement of the form

$$c = \iota v\ \varphi,$$

where: c is a constant of the suitable sort; ι is the *iota operator* of definite descriptions; v is a variable of the same sort of c; and φ is a formula containing v as the only free variable. As before, I assume that answers to wh-questions must be given within k instants:

(AskWh) $askWh(x,y,\varphi) = request(x,y,Done(y,assert(y,x,c = \iota v\ \varphi)/now+k))$,

where v is the only variable free in φ, and c is a constant.

3.9 Proposals

A proposal is the conjunction of a directive and a conditional commissive. For example, we can analyze "x proposes to y to buy item i" in the following terms:

- x precommits y, relative to x, to transferring the property of i to x; and
- x commits, relative to y, to transferring a given amount of money to y, on condition that y transfers the property of i to x or commits to doing so.

Indeed, proposals are going to be very common in agent interactions. We can expect them to show up every time an elementary interaction between two agents, x and y, involves the execution of an action by x and of another action by y. This is the case, for example, with commercial transactions, where the same event can be described as *buying* from the point of view of a client, and as *selling* from the point of view of a seller.

To deal properly with proposals, let us first introduce a way to treat elementary interactions. In an elementary interaction, we have a number of *participants*, each of which has a specified *role*. In turn, a role is an action or a set of actions, performed by one of the participants. For example, buying item i at price p can be defined as:

(Buy1) $Role(x,buy(x,y,i,p),moneyTransfer(x,y,p))$,

(Buy2) $Role(y,buy(x,y,i,p),propertyTransfer(y,x,i))$.

Proposals can be defined as follows:

(Propose1) $propose(x,y,\varphi) = speechAct(x,y,propose,\varphi)$,

(Propose2) $Done(e,x,propose(x,y,Done(x,d,\alpha))) \wedge Role(x,\alpha,\beta) \wedge Role(y,\alpha,\gamma) \wedge d > now \rightarrow PC(e,y,x,Done(y,d,\gamma)) \wedge (Done(y,d,\gamma) \vee C(e,y,x,Done(y,d,\gamma)) \rightarrow C(e,x,y,Done(x,d,\beta)))$.

This definition has the following consequences. First, as a proposal by x to y puts y in a state of precommitment, it is meaningful for y to accept a proposal. Second, if y accepts the proposal, both agents are committed to their respective roles in the interaction; x's commitment is brought about also if y directly performs its role without explicitly committing to it.

4. COMMITMENTS, NORMATIVE SYSTEMS, AND RATIONALITY

Communication is the interface between the internal, mental world of an agent and the external world of social interaction. Therefore, it is important to connect the semantics of speech acts to the rationality principles governing individual action and to the normative systems that regulate social interactions.

In the traditional approach to ACL semantics, one attempts to define the meaning of speech acts in terms of mental states, and tends to view semantic conditions as special cases of general rationality principles (see for example [5]). The connection between communication and individual rationality is therefore implicit in the semantics of speech acts. In this paper, I have

followed an alternative approach, regarding speech acts as actions involving social effects. To which extent an agent should act rationally is, in my opinion, an important matter for agent design, but should not be part of the semantics of a communication language [7]. It is however interesting to find out principled ways of interfacing the semantics of speech acts with the world of individual rationality. My suggestion is that such a connection can be realized through the concept of *violation*.

It is part of the very nature of commitments that they can be violated. As a general idea, a commitment is violated if its content does not hold. We can therefore define a violation of the commitments of agent x relative to agent y created by action e as follows:

$$V(e,x,y) =_{\text{def}} C(e,x,y,\varphi) \wedge \neg\varphi.$$

The concept of violation allows one to connect commitments to both social institutions and individual rationality. First, in a society of agents violations might be discovered and recorded by some independent authority (itself an agent), which might also be in charge of establishing sanctions for the offending agents. It seems reasonable to assume that the same kind of violation can lead to different sanctions depending on the *normative system* regulating a specific type of interaction. For example, the consequences of a false assertion can be very different if the assertion is made in an informal conversation or as part of a commercial transaction. To avoid ambiguities, agents will have to make it explicit under which normative system they are carrying on a conversation.

A second, important role of the concept of violation is to allow an agent to reason on commitments in order to plan a rational course of action. Suppose for example that, as a result of action e_1, agent a is committed, relative to agent b, to transferring the property of item i to b:

(C1) $C(e_1,a,b,propertyTransfer(a,b,i))$.

Intuitively, we would like to derive from C1 that a is committed not to transfer the property of i to any agent x different from b, but the logic I have adopted for commitment is so weak that this derivation cannot be carried out. There is, however, an alternative solution. From general knowledge on property transfer, an agent may still infer that, given C1, transferring the property of i to an agent different from b logically entails a violation of the commitments created by e_1. In a natural deduction style, the derivation would look like this:

1. $C(e_1,a,b,propertyTransfer(a,b,i))$	premise,
2. $b \neq c$	premise,
3. $Done(e,propertyTransfer(a,c,i))$	premise,
4. $\neg Done(e',propertyTransfer(a,b,i))$	from 2, 3, and a theorem on property transfer,
5. $V(e_1,a,b)$	from 1, 4, and the definition of violation.

This result is sufficient for a to plan a course of action that does not bring about a violation. Of course, a may also decide to transfer the property of i to an agent different from b, thus violating commitment C1, if the reward expected from such an action exceeds the sanction expected for the violation. Such a behavior may well be considered as ethically questionable, but it is economically rational, at least on a short term. To decide whether an agent should actually behave in such a way is then up to agent designers.

My working hypothesis is that reasoning on violations can completely compensate for the weakness of the logic of commitment. In particular, nothing forbids an agent to make conflicting commitments, because in my logic of commitment there is no counterpart of the D axiom of deontic logic. However, general inferential capacities are sufficient to derive that at least one of two conflicting commitments is going to be violated, and that this violation can be expected to bring about a sanction. An agent may therefore avoid making conflicting commitments on the purely economical ground that violations are costly.

Reasoning on violations also allows us to deal with *conditional commitments*. In deontic logic, conditional obligations are a remarkable source of troubles. The following example shows that the same kind of difficulties arise also with commitment. Suppose that, as a result of action e_1, agent a commits, relative to b, to the fact that q holds, under the condition that p holds. We have two ways of expressing this in the logical language. The first is to state a commitment with a conditional content:

(C2) $C(e_1,a,b,p \rightarrow q)$.

The second is to state a conditional formula with p as the antecedent, and a commitment with content q as the consequent:

(C3) $p \rightarrow C(e_1,a,b,q)$.

Both solutions have pros and cons. However, if we regard the conditional commitment as the result of an assertion, we are forced to choose C2: if the conditional statement '$p \rightarrow q$' is the content of an assertion, there is no natural way to end up with a commitment having the form of C3. Now the question is, What happens if both C2 and p hold? Intuitively, we would like the following statement to derive from C2 and p:

(C4) $C(e_1,a,b,q)$.

But C4 does not derive from C2 and p, and would not do so even if we adopted a much stronger logic of commitment (like, for example, a normal modal logic including axiom K). This difficulty is well known in deontic logic, and has lead many logicians to base their systems on a primitive binary operator of conditional obligation (see [2]) – a solution, however, that has a number of problems of its own.

It is interesting to see that reasoning on violations provides a straightforward solution to the treatment of conditional commitments. Consider again sentence C2, and assume that p holds. Even if a commitment to q cannot be explicitly derived, it is possible to infer that $\neg q$ is going to violate the commitments created by e_1. Here is a sketch of a possible derivation:

1.	$C(e_1,a,b,p \rightarrow q)$	premise,
2.	p	premise,
3.	$\neg q$	premise,
4.	$\neg(p \rightarrow q)$	from 2 and 3,
5.	$V(e_1,a,b)$	from 1, 4, and the definition of violation.

As in the case of conflicting commitments, this is sufficient for a to plan a rational course of action.

5. CONVERSATIONS

The ultimate goal of an ACL is to allow agents to carry on *conversations*. Conversations are sequences of turns, and a turn is a single speech act or a set of contiguous speech acts performed by a single agent.

The problem with conversations is to guarantee that they are carried on coherently. Broadly speaking, there are two standard solutions. The one traditional in AI is to regard conversations as special cases of *rational interaction* (see for example [12]). Conversations are assumed to emerge as an effect of two or more agents trying to pursue their individual goals on the basis of their beliefs about each other. A diametrically different approach is based on the definition of *conversational protocols*, which constitute a sort of dialogue grammar dictating what sequences of speech acts are to be considered as well formed. The scheme of *conversation for action* proposed by Winograd and Flores [20] is a paradigmatic example of this approach.

In my opinion, the rational interaction model is not practically viable. All social interactions, including conversations, do embody a significant amount of rationality, but this does not imply that agents should generate their interactions through an on-line process of practical reasoning. Such a process would be highly inefficient and beyond the capabilities of simple, reactive agents.

Conversational protocols do not suffer from this difficulty: they can be implemented efficiently and do not require an agent to have a complex, deliberative architecture. Also with this approach, however, there are a couple of problems. The first difficulty is theoretical. What should be the relationship between conversational protocols and speech act semantics? Do conversational protocols add something to the semantics of speech acts? Or are they second-level structures, which have to be compatible with speech acts semantics, but do not add anything to it? The second problem is practical. If agents are allowed to use an extensible ACL, how are newly defined speech acts to be dealt with in conversations? Does the introduction of a new type of speech acts impose the definition of new conversational protocols? If so, do we have well-defined adequacy criteria for the introduction of such protocols?

To solve these problems, I think there are good reasons to adopt the following standpoint:

(i) *The semantics of messages should be independent of the structure of conversations, and completely defined at the level of speech acts.* In fact, if we accept semantics to be partially defined at the level of conversations, we cannot assign an unambiguous meaning to a speech act type. It seems to me that a similar situation would bring in severe problems at the level of agent specification and design.

(ii) *The structure of well-formed conversations should be derived from general conversational principles.* The reason is that we need general criteria of what counts as a coherent conversation *before* we define conversational protocols.

(iii) *Conversational protocols are specific implementations of general conversational principles.* They can be of great practical utility, but do not contribute to the semantics of messages and, at least in principle, should be derivable from general principles of conversation.

My definition of speech acts meets criterion (i), because the semantics of messages is completely defined in terms of commitments and precommitments, and does not refer to conversations. Criterion (ii) is less trivial: what kinds of conversational principles can be practically adopted? My working hypothesis is that also conversational principles can be cast in terms of commitments, in particular of *conversational commitments*.

Let me try to clarify this concept with an example. Suppose that agent a requests agent b to perform some action like, for instance, switching the microwave oven off within the next ten minutes. As a result of this speech act, b is precommitted to switching the microwave oven off within the next ten minutes. We now expect b to react to a's directive in the appropriate way, that is, by switching off the oven immediately, or by accepting the precommitment, or by rejecting it (in a more complex context, b may have further options, like negotiating a different time limit). On the contrary, we are not ready to accept that b just keeps silent. The principle that seems to be at work is that an agent engaged in a conversation has to take an explicit stand with respect to all its precommitments. The basic options are to directly fulfill the precommitment, to accept it, or to reject it.

Casting conversational rules in terms of reactions to precommitments places such rules at an intermediate level between general rationality principles and conversational protocols. The main advantage over protocols is that one does not need to take into account the whole range of speech acts that can be realized in a given language. A conversational rule defined in term of reactions to precommitments will be triggered every time a precommitment is created, and will not depend on the specific speech act that has created it.

As we have already seen, precommitments to do an action come with a definite time limit, d, within which the action should be performed. I now need to introduce a second time limit, $d' \leq d$, within which an agent is required to explicitly react to its precommitments. The following axiom states that any ordinary precommitment comes with an associated conversational precommitment:

(CPC) $PC(e,x,y,\varphi) \rightarrow PC(e,x,y,Done(x,react(e)/d'))$.

Note that CPC does not create an infinite sequence of distinct precommitments: the only precommitments it generates has the form $PC(e,x,y,Done(x,react(e)/d'))$. The reaction time limit, d', may be treated as a global constant, or as a further parameter of all speech acts that create precommitments. The term $react(e)$ denotes a new action type, defined as follows:

$$PC(e,x,y,Done(x,d,\alpha)) \rightarrow (Done(x,react(e)/d') \leftrightarrow Within(d',Done(x,\alpha) \vee C(e,x,y,Done(x,\alpha/d)) \vee \neg PC(e,x,y,Done(x,\alpha/d))).$$

The two axioms above define the conversational precommitment implicitly created by ordinary precommitments. So far, however, no conversational commitment is generated. To do so is up to *conversational contracts*. For example, we may specify a *standard conversational contract* stating that all precommitments introduced by CPC are directly turned into commitments. The formal definition of such a contract is straightforward:

$$Contract(x,y,react(e)).$$

In particular contexts, however, one may want to define different conversational contracts. Consider for example a situation in which an agent, a, is bound by an ordinary contract to accept all precommitments created by another agent, b. In such a case, the standard conversational contract may be considered redundant, and can be dropped.

Let me remark that I regard conversational axioms like the ones presented above as specifications of coherent conversations, not as protocols. It is up to designers to implement the conversational competence of their agents as on-line mechanisms of deduction from general conversational axioms or as pre-compiled protocols. In the latter case, however, the conversational axioms allow one to check whether a protocol correctly implements the general principles of conversation that have been adopted.

The model I have developed so far is clearly incomplete. Intuitively, there is much more to conversations than the reaction to ordinary precommitments. However, my working hypothesis, which will have to be put to test through theoretical development and experimentation, is that all rules of conversation can be cast into the form of conversational precommitments. Here are some examples:

- an agent's failure to decode the body of a message creates a conversational precommitment to assert that the message has not been understood;
- an agent's failure to fulfill a commitment creates a conversational precommitment to notify the failure to the commitment's creditor;
- fulfilling a commitment creates a conversational precommitment to notify the fulfillment to the commitment's creditor.

6. CONCLUSIONS

In this paper I have sketched a model of agent communication. The main feature of my proposal is the commitment-based semantics of speech acts. Other potentially important aspects are, for the moment, at the level of working hypotheses. Among these I regard: (i) the use of violations to build logical connections among commitments, external normative systems, and individual rationality; and (ii), the idea that conversations may be dealt with in terms of conversational precommitments and contracts.

The level at which communication is treated in this paper is very abstract, and there is a considerable gap to fill in order to bring the model down to the level of implementation. My personal view is that most of the agents used in practical applications will not have a high degree of "intelligence" (i.e., on-line inferential capacities). Therefore, I regard the model proposed mainly as a means to specify communicating agents off line. However, much work is still needed before my proposal can be turned into a usable tool, and some of this work will necessarily be experimental. My current plans are to apply the model to trading agents (in particular in the contexts of auctions) and to a community of agents in charge of managing an intelligent building.

ACKNOWLEDGMENTS

Most of the ideas contained in this paper have been presented in seminar form at Politecnico di Milano, Università della Svizzera italiana, and Università Cattolica "Sacro Cuore". I am grateful to several colleagues of these universities for useful criticisms and comments. I thank Alessio Lomuscio for many useful email exchanges and for commenting upon a draft version of the paper, and Claudio Signorini for interesting discussions and for implementing a preliminary version of the model as part of his Master Dissertation in Computer Engineering at Politecnico di Milano (1999).

REFERENCES

[1] Airenti, G., B.G. Bara and M. Colombetti (1993). Conversation and behavior games in the pragmatics of dialogue. *Cognitive Science* 17, 197–256.

[2] Åqvist, L. (1984). Deontic Logic. In D. Gabbay and F. Guenthner, eds., *Handbook of philosophical logic* II, Reidel, Dordrecht, D, 605–714.

[3] Austin, J.L. (1962). *How to do things with words*, Clarendon Press, Oxford, UK.

[4] Chellas, B.F. (1980). *Modal logic*, Cambridge University Press, Cambridge, UK.

[5] Cohen, P.R., and H.J. Levesque (1990). Rational interaction as the basis for communication. In P.R. Cohen, J. Morgan and M.E. Pollack, eds., *Intentions in communication*, MIT Press, Cambridge, MA, 221–256.

[6] Cohen, P.R., and H.J. Levesque (1995). Communicative actions for artificial agents. *Proceedings of the International Conference on Multi-Agent Systems*, AAAI Press, San Francisco, June 1995.

[7] Colombetti, M. (1999). Semantic, normative and practical aspects of agent communication. *Preprints of the IJCAI'99 Workshop on Agent Communication Languages*, Stockholm, 51–62.

[8] Finin, T., Y. Labrou, and J. Mayfield (1995). KQML as an agent communication language. In J. Bradshaw, ed., *Sofware agents*, MIT Press, Cambridge, MA.

[9] FIPA (1997). Agent Communication Language. FIPA 97 Specification, Foundation for Intelligent Physical Agents, http://www.fipa.org.

[10] Jones, A.J.I., and M.J. Sergot (1996). A formal characterisation of institutionalised power. *Journal of the IGPL* 4, 429–445.

[11] Labrou, Y., and T. Finin (1997). Semantics and conversations for an agent communication language. *Proceedings of the 15th International Joint Conference on Artificial Intelligence* (*IJCAI'97*), Nagoya, Japan.

[12] Litman, D.J., and J.F. Allen (1990). Discourse processing and commonsense plans. In P. R. Cohen, J. Morgan and M. E. Pollack, eds., *Intentions in communication*, MIT Press, Cambridge, MA.

[13] Pitt, J., and A. Mamdani (1999). A protocol based semantics for an agent communication language. *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI'99)*, Stockholm, 486–491.

[14] Searle, J.R. (1969). *Speech Acts*, Cambridge University Press, Cambridge, UK.

[15] Searle, J.R. (1975). A taxonomy of illocutionary acts. In K. Gunderson, ed., *Language, mind, and knowledge* (*Minnesota Studies in the Philosophy of Science* VII), University of Minnesota Press, 344–369. Reprinted in J. R. Searle (1979), *Expression and meaning*, Cambridge University Press, Cambridge, UK.

[16] Searle, J.R., and D. Vanderveken (1985). *Foundations of illocutionary logic*, Cambridge University Press, Cambridge, UK.

[17] Singh, M.P. (1998). Agent communication languages: Rethinking the principles. *IEEE Computer* 31, 40–47.

[18] Singh, M.P. (1999). A social semantics for agent communication languages. *Preprints of the IJCAI'99 Workshop on Agent Communication Languages*, Stockholm, 75–88.

[19] Singh, M.P. (2000). Synthesizing coordination requirements for heterogeneous autonomous agents. *Autonomous Agents and Multi-Agent Systems*, 3, in press.

[20] Winograd, T., and F. Flores (1986). *Understanding computers and cognition: A new foundation for design*, Ablex, Norwood, NJ.