An Analysis of Agent Speech Acts as Institutional Actions

Marco Colombetti
Politecnico di Milano
Piazza L. Da Vinci, 32, 20133 Milano, Italy
University of Lugano
Via Buffi 13, 6900 Lugano, Switzerland
Marco.Colombetti@polimi.it

Mario Verdicchio
Politecnico di Milano
Department of Electronics and Information
Piazza L. Da Vinci, 32
20133 Milano, Italy
Mario.Verdicchio@elet.polimi.it

ABSTRACT

All major proposals in the field of Agent Communication Languages deal with agent communication in terms of speech acts. This choice is important not only because it allows one to rely on a powerful and deep theory of communication, namely Speech Act Theory, but also because AI has developed computationally effective ways of dealing with actions. However, the AI literature does not seem to distinguish between "physical" or "natural" actions and speech acts in a principled way. This attitude often results in a fairly confused and inadequate account of what it means that an agent performs a speech act – a situation that is likely to hinder further developments in the field of Agent Communication Languages. In this paper we analyze the concept of speech act, and point out the main differences between speech acts, conceived as a special category of "institutional" actions, and natural actions. On the basis of our analysis, we conclude that speech acts should be modelled in terms of the specific social effects brought about by their performance.

Categories and Subject Descriptors

I.2.0 [**Artificial Intelligence**]: General—*Philosophical foundations*; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent systems*

Keywords

Agent Communication Language, Speech Act, Action, Commitment

1. INTRODUCTION

Agent Communication Languages (ACLs) have recently become a primary concern in the field of multiagent systems, especially in connection with open systems, which are generally considered to need a standard communication framework shared by all interacting agents. However, the task of establishing a successful language standard is problematic because of a number of issues. A standard should be completely and rigorously defined, yet flexible and extendible, to let agents cope with varied and new undertakings. Moreover, an ACL has to be simple enough to be correctly understood and used by agent designers, yet enough expressive to allow for every significant kind of agent conversation to be carried out. If none of the proposals put forward so far has been universally accepted, it may be because the abovementioned issues have not tackled in a satisfactory way.

Even if we do not yet have a universally accepted standard ACL, it is interesting to note that all major proposals share the assumption that agent communication should be dealt with in terms of speech acts. As is well known, the notion of a speech act comes from philosophy of language [1, 14]. After playing an important role in AI models of human-machine communication, starting from a pioneering paper by Cohen and Perrault [3], speech acts have been adopted as the base of agent communication by the proposers of KQML [8] and by the Foundation for Intelligent Physical Agents [9]. Notably, a speech-act based view of agent communication is adopted also by scientists, like Singh [19], whose approach is radically different from the one underlying KQML and FIPA ACL.

The reasons why the concept of speech act has been so successful are not difficult to understand. Firstly, Speech Act Theory, as worked out by Austin and Searle, is a powerful, deep and remarkably comprehensive theory of language semantics. Besides, as the development of agents that plan and act rationally is one of AI's strongholds, regarding communication as a form of action makes it easy and natural to embed a communicative component in an agent's global architecture. Finally, even if they are named after it, speech acts need not have anything to do with speech in a literal sense. On the contrary, they can be realized in any symbolic system, either linguistic or non-linguistic. In particular, this makes it possible to treat multi-modal human-computer interaction in terms of speech acts.

Speech acts have been adopted so promptly by computer scientists also because AI has developed computationally effective ways of dealing with actions. Formal theories of action have been extensively studied since the invention of Situation Calculus [11]. On the applications' side, however, the most influential model is the one pioneered by STRIPS [7], presented by several introductory books as "the" AI model of action, and typically exemplified on the blocks world [12, 13]. This model, suitable for simple physical actions, is so easy to understand and use that it has deeply influenced the treatment of action in general, including speech acts. In

AAMAS'02, July 15-19, 2002, Bologna, Italy.

fact, the AI literature does not seem to distinguish between "physical" or "natural" actions and speech acts in a principled way. This attitude results, in our opinion, in a fairly confused and inadequate account of what it means that an agent performs a speech act – a situation that is likely to hinder further developments in the field of ACLs.

In this paper we want to analyze the concept of speech act, in hope of shedding some light on the fundamental differences between speech acts and natural actions. Such differences, we believe, should in turn be reflected into different formal treatments. To carry out our analysis we shall keep the following route. First (Section 2) we propose a definition of action that allows us to distinguish between primary and secondary natural actions. Then (Section 3) we define the concept of institutional action, and argue that speech acts (more precisely, utterance and illocutionary acts) are special kinds of institutional actions. In Section 4 we criticize a current approach to the definition of speech acts, and defend a different proposal. Finally, we draw some conclusions in Section 5.

2. EVENTS AND ACTIONS

Intuitively, an action is an event intentionally brought about by an agent. In this section we briefly analyze this intuition, largely relying on Searle's philosophical analysis [16]. To achieve clarity and rigour, we use a logic-like notation. However, we do not propose here a full-fledged formal account of action. In particular, we have not yet developed a formal semantics for our notation, nor a systematic treatment of temporal aspects.

2.1 Events

We reify events, that is, we treat events as a kind of individuals, called *event tokens*. Every event token belongs to (at least) an *event type*, and we write $Event(e,t)$ to say that e is a token of type t.

An event type can be defined in terms of a change in the state of the world, understood as the set of all properties of objects and relationships among them. We assume that every event token occurs over a closed interval of time (conceived as a discrete, dense or continuous linear order of time instants), including the limiting case of a time point. We adopt the following notation:

- $\phi[e$ means that formula ϕ is true at the starting point of e's interval;
- $e]\phi$ means that ϕ is true at the ending point of e's interval;
- $\phi\langle e$ means that ϕ is true immediately before the starting point of e's interval;
- $e\rangle\phi$ means that ϕ is true immediately after the ending point of e's interval;
- $e : \phi$ means that ϕ is true at all internal points of e's interval.

Suppose for example that, within a suitable logical theory of objects in space, the atomic formula $At(o,l)$ means that object o is at location l. Here are some sample definitions of event types:

- reaching a location:

$$Event(e, reach(o,l)) \triangleq \neg At(o,l)\langle e]At(o,l);$$

- leaving a location:

$$Event(e, leave(o,l)) \triangleq At(o,l)[e\rangle\neg At(o,l);$$

- moving from a location to another one:

$$Event(e, move(o,l',l'')) \triangleq At(o,l')[e : \neg At(o,l') \wedge \neg At(o,l'')]At(o,l'').$$

The above definitions have a number of interesting features. First, they are analytical, in that they provide necessary and sufficient conditions for an event of the given type to take place. Second, they are given solely in terms of world states. Third, they say nothing about the particular process or procedure by which the event is brought about.

Event tokens may be related by causal links. When event e causes event e' we write $Cause(e,e')$. For example, here is a statement that object o_1 reaching location l_1 causes object o_2 to leave the same location:

$$Event(e_1, reach(o_1,l_1)) \wedge Event(e_2, leave(o_2,l_1)) \wedge Cause(e_1,e_2).$$

The following axiom captures a fundamental property of causation:

(A1) $\quad Cause(e,e') \wedge Cause(e',e'') \rightarrow Cause(e,e'').$

2.2 Actions

As we have already said, an action is an event intentionally brought about by an agent. Agents have two characteristic features. First, they can entertain mental states, like beliefs, intentions, and desires. Second, for each agent there is a set of event types, which we shall call *primary actions*, whose tokens can be intentionally brought about by the agent without intentionally bringing about any other event token. The primary actions that an agent may execute depend on the agent's basic abilities; therefore, distinct agents may have a different repertoire of primary actions. For a human being, a primary action may be opening an eye or stretching a finger. For a robot, a primary action is one that is directly performed by running the driver of an effector.

While primary actions do not need the execution of any other action to be carried out, *secondary actions* can only be performed through the execution of another action. For example, a human agent may turn on a computer by pressing a button: turning on the computer is here a secondary action. To press the button, the agent will have to perform certain bodily movements, and thus also pressing the button is a secondary action. On the contrary, the bodily movements required to press the button are typically primary actions.

2.2.1 Primary actions

To state that e is a token of a primary action of type t, intentionally performed by agent a, we write $Action_1(a,e,t)$. This amounts to saying that:

(1.1) type t is in the repertoire of a's primary actions;

(1.2) a intends to bring about an event of type t;

(1.3) such an intention causes e to occur;

(1.4) event e has type t.

We already know how to express Clauses 1.3 and 1.4 formally. As far as Clause 1.1 is concerned, we simply write $PrimaryType(a,t)$ to state that t is a primary action type for agent a. The most critical point is Clause 1.2, as it involves a formal representation of intentions. First let us observe that the concept we need here is what Searle [16] calls intention-in-action, that is, the distinctive mental state that agents entertain during the intentional execution of an action. (Intentions-in-action should not be confused with *prior intentions*, which are the building blocks of action plans to be executed in the future.) The main feature of an intention-in-action, according to Searle, is that it both represents and directly causes the action that is carried out under its control.

If intentions are to cause events, it seems reasonable to treat them as a special category of events. We shall therefore write $Event(e, intend_1(a,t))$ to state that e is (a token of) an intention by a to perform a primary action of type t. Among all categories of events, mental events are special, in that they have propositional content. This means that mental events have conditions of satisfaction, that is, conditions under which their propositional content holds. We plan to formalize this feature of mental events within a possible world approach; however, we expect that capturing the relationship between the representational and the causal components of intentions may prove difficult.

Clauses 1.1–1.4 can now be formalized as:

$$
\begin{aligned}
&Action_1(a,e,t) \triangleq \\
&\quad PrimaryType(a,t) \wedge && (1.1)\\
&\quad \exists e'(Event(e', intend_1(a,t)) \wedge && (1.2)\\
&\qquad Cause(e',e)) \wedge && (1.3)\\
&\quad Event(e,t). && (1.4)
\end{aligned}
$$

2.2.2 Secondary actions

A secondary action is an event intentionally brought about through the execution of another action. We write

$$Action_2(a,e,t,e',t')$$

to state that e is a token of a secondary action of type t, intentionally performed by agent a through the execution of another action e' of type t'. This amounts to saying that:

(2.1) e' is an event of type t';

(2.2) by bringing about an event of type t', a intends to bring about an event of type t;

(2.3) such an intention causes e' to occur;

(2.4) e is an event of type t;

(2.5) e' causes e to occur.

Clauses 2.1 and 2.4 are formally stated as

$$Event(e,t) \wedge Event(e',t').$$

To state Clause 2.2 we need to represent the intention to bring about an event of type t by bringing about an event of type t'. We represent this as a mental event:

$$Event(e'', intend_2(a,t,t')).$$

Our definition of secondary action then becomes

$$
\begin{aligned}
&Action_2(a,e,t,e',t') \triangleq \\
&\quad Event(e',t') \wedge && (2.1)\\
&\quad \exists e''(Event(e'', intend_2(a,t,t')) \wedge && (2.2)\\
&\qquad Cause(e'',e')) \wedge && (2.3)\\
&\quad Event(e,t) \wedge && (2.4)\\
&\quad Cause(e',e). && (2.5)
\end{aligned}
$$

It is important to remark that the two function symbols $intend_1$ (used in subsection 2.2.1) and $intend_2$ (used in this subsection) are distinct. In particular, they differ in the number of arguments. From a strictly formal point of view, it would be possible to use only one symbol, $intent_2$, by defining $intend_1(a,t)$ to be the same as $intend_2(a,t,nil)$, where nil denotes an "empty action type." However, we are inclined to think that further formal analysis will show that the difference between the two kind of intentions is not superficial, but substantial. Therefore, we prefer to denote the two concepts by two distinct symbols.

2.2.3 Success and failure

The definitions given so far describe successful actions, that is, actions whose execution satisfies the corresponding intention. But what about failures?

A failed action is one whose execution does not satisfy the corresponding intention. But this is not enough. To fail, an action must at least be attempted. So, in order to define failure we must define what it means to attempt an action. Let us start with primary actions: we want to define $Attempt_1(a,t)$ to mean that agent a attempts a primary action of type t. To do so it is sufficient to extract Clauses 1.1 and 1.2 from the definition of $Action_1$:

$$
\begin{aligned}
&Attempt_1(a,t) \triangleq \\
&\quad PrimaryType(a,t) \wedge && (1.1)\\
&\quad \exists e'(Event(e', intend_1(a,t)). && (1.2)
\end{aligned}
$$

The situation is similar with secondary actions. In this case, we want to define $Attempt_2(a,t,e',t')$ to mean that agent a performs action e' of type t' to attempt a secondary action of type t. Here it is sufficient to take Clauses 2.1–2.3 from the definition of $Action_2$:

$$
\begin{aligned}
&Attempt_2(a,e,t,e',t') \triangleq \\
&\quad Event(e',t') \wedge && (2.1)\\
&\quad \exists e''(Event(e'', intend_2(a,t,t')) \wedge && (2.2)\\
&\qquad Cause(e'',e')). && (2.3)
\end{aligned}
$$

2.3 The Relation with AI Representations

It is important to understand the relationships between our definition and the traditional representations of actions in AI systems. Suppose for example we want to define the action of switching on a computer by pressing the ON/OFF button. A typical AI representation would go like this:

action *switchOn*(*x*:*computer*)
  preconditions *Off*(*x*), *PluggedIn*(*x*)
  add *On*(*x*)
  delete *Off*(*x*)
  do *pushButton*(*x*)
end.

Similar definitions merge different aspects together. First, the analytical definition of the event type "switching on" is provided by precondition $Off(x)$ together with the add and delete clauses. In our notation:

$$Event(e, switchOn(x)) \triangleq \neg On(x) \langle e \rangle On(x).$$

Second, the procedure $pushButton(x)$ and the precondition $PluggedIn(x)$ specify the instrumental action by which an agent is expected to switch on a computer, and a necessary condition for this instrumental action to cause the computer switching on.

It seems to us that, in general, AI representations of actions tend to mix together the analytical definition of an event type, the description of an instrumental action that may cause the target event, and a set of necessary (and hopefully sufficient) conditions for the instrumental action to cause the target event. This representation scheme has proved satisfactory in many applications involving natural actions but, as we shall try to show, is not adequate to represent other kinds of actions, and in particular speech acts.

3. INSTITUTIONAL EVENTS AND ACTIONS

In Section 2 we have sketched a definition of action that appears to be fit for natural actions, that is, for actions that only involve an agent's physical abilities and processes of physical causation. This definition is adequate to deal with agents acting individually in a physical environment, but does not account for *institutional actions*, like playing games, buying and selling goods, or performing speech acts [17].

3.1 Institutional events

We start our analysis from institutional events. To understand this concept better, let us take as an example an event of property transfer from an agent to another one (as it may occur in a donation or as part of a commercial transaction). Let us suppose that, within a logical theory of property, formula $Prop(a, o)$ means that agent a is the proprietor of object o. Then an event of property transfer may be defined as follows:

$$Event(e, propTrans(a, b, o)) \triangleq Prop(a, o)\langle e\rangle Prop(b, o).$$

Like in the examples of Section 2, this formula defines an event type analytically, it is given in terms of states, and says nothing about the procedures that can bring about a token of the event type. With respect to natural events, however, there are two important differences. The first difference is that the properties involved in the definition of an institutional event are not natural, but are themselves institutional, like in the preceding example on property. As we shall see in the following, this fact has important consequences on the distinction between perlocutionary and illocutionary speech acts and on the definition of illocutionary acts. The second difference is that, contrary to natural events, an institutional event is not brought about by exploiting causal links. As argued by Searle [17], institutional events are realized through the "counts as" relationship. That is, an institutional event of type t is brought about by an event of type t', that counts as an event of type t in an appropriate context. Which events may count as cases of property transfer is established by the social institutions that regulate property. For example, the social institutions of many human societies establish that a's handing on to b an object at b's birthday party while saying "Happy birthday!" counts as a case of property transfer.

Now take $CountsAs(e, t, t')$ to mean that event e of type t counts as an event of type t' according to some institution. We state the following axiom:

$$\text{(A2)} \quad Event(e, t) \wedge CountsAs(e, t, t') \rightarrow Event(e, t').$$

We have now to say how institutions can be represented. As we have already remarked, an event of type t' typically counts as an event of type t only in an appropriate context. This means that, as far as the "counts as" relation is concerned, an institution is represented by a logical theory establishing under which contextual conditions the formula $CountsAs(e, t, t')$ holds.

A fully formal analysis of human institutions in these terms would be a huge enterprise - maybe an impossible one. But this need not be the case with artificial agents. On the contrary, a formal definition of the contextual conditions under which certain institutional acts can be performed by an artificial agent is going to be necessary for the development of agent-based technology in many fields, like for example electronic commerce.

3.2 Institutional actions

As we have seen in Section 2, to perform secondary actions agents exploit causal links. Analogously, an agent may exploit a "counts as" relation to realize an institutional action. We now define what it means to perform an institutional action of type t' by means of an action of type t (either natural or institutional). As we did in Section 2, we start by representing the corresponding intention, that is, the intention to perform an action of type t by performing an action of type t' that counts as an action of type t (according to some institution): $Event(e, intend_I(a, t, t'))$. We now define the performance of an institutional action as:

$$\begin{aligned} &Action_I(a, e, t, t') \triangleq \\ &\quad Event(e, t) \wedge \\ &\quad \exists e'(Event(e', intend_I(a, t, t')) \wedge \\ &\qquad Cause(e', e)) \wedge \\ &\quad CountsAs(e, t, t'). \end{aligned}$$

Note that for this definition to make sense it is necessary that t be an institutional event type.

We are now able to define action in general (either primary, secondary, or institutional) as:

$$\begin{aligned} &Action(a, e, t) \triangleq \\ &\quad Action_1(a, e, t) \vee \\ &\quad \exists e' \exists t' Action_2(a, e, t, e', t') \vee \\ &\quad \exists t' Action_I(a, e, t, t'). \end{aligned}$$

The logical structure of institutional actions is sharply different from that of natural actions. First, there cannot be any primary institutional action: an agent's abilities, by themselves, cannot bring about an institutional event. Second, a secondary action involves a causal link between two distinct and independently defined events (like "pushing the button" and "switching on the computer"), that may separately occur even if there is no causal link between them. On the contrary, when an institutional action of type t is performed as an action of type t' that counts as an event of type t, only one event token is involved, whose type is promoted, so to speak, from t' to t by the "counts as" relation (see Fig. 1).

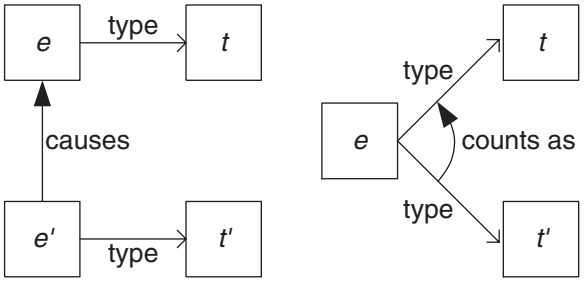

Figure 1: The "causes" and "counts as" relationships.

3.3 Speech acts

Speech acts are a special category of institutional acts [14]. In the rest of this paper we shall deal with two kinds of speech acts, namely utterance acts and illocutionary acts. We shall also say a few words about perlocutionary acts, in order to point out what we believe to be a common misunderstanding.

3.3.1 Utterance acts

As an event type, an utterance involves a sentence s in some language, uttered by an agent, a, and addressed to another agent, b. To say that e is a token of such a type we write $Event(e, utter(a, b, s))$.

Utterances are institutional events, because what is a sentence of a language is established by the grammar of such a language (a linguistic institution). An agent performs an utterance by executing a (typically primary) action that counts as an utterance according to the grammar of some language. Typically, human agents do so by saying something, that is, by producing certain sound patterns with their phonatory organs. An artificial agent may perform an utterance by sending a message to another agent. In all cases, the key point is that some institution guarantees that, in an appropriate context, the primary action counts as an utterance act.

3.3.2 Illocutionary acts

Under given conditions, an utterance act counts as an illocutionary act. For example, saying "My name is Philip" typically counts as an assertion, which is a type of illocutionary act.

According to the most widely accepted theory of speech acts [14, 15, 18], illocutionary acts are classified in five types: assertives, commissives, directives, expressives, and declarations. An utterance act counts as an illocutionary act of a given type on the basis of certain features of the sentence uttered and of a number of contextual conditions, which are characteristic of the type.

Even if we can rely on a general and deep philosophical theory of illocution, building a formal model of the illocutionary acts performed by artificial agents has proved a difficult task. In Section 4 we briefly analyze an existing proposal and sketch the guidelines for a different one.

3.3.3 Perlocutionary acts

An illocutionary act is performed by an agent to achieve certain effects on another agent. For example, assertives are typically performed to convince the addressee that some state of affairs holds, and directives are typically performed to induce the addressee to carry out some action.

Contrary to utterance and illocutionary acts, perlocutionary acts are not institutional actions. In fact, no action performed by agent a can "count as" an action of convincing agent b about something, or of inducing b to do an action. To do so, a has to actually cause a suitable mental state of b's: a belief in the case of convincing, and an intention in the case of inducing. The relation between illocution and perlocution is one of causation; perlocution is therefore a secondary action, achieved by the execution of an illocutionary act. It may be objected that these cases of mental causation cannot be treated as the cases of physical causation underlying natural actions. The main difference is due to agents' autonomy. In fact, an autonomous agent cannot be forced to have a given mental state; rather, it has to be provided with *reasons* to entertain a new belief, desire, or intention. But, even if the mechanisms called for are different from those involved in physical causation, an illocutionary act that achieves a perlocutionary effect may well be said to cause the relevant mental states of the addressee. We conclude that perlocutionary acts are natural, secondary actions, performed by the execution of an institutional action (an illocutionary act).

The spectrum of realization relations among actions is depicted in Figure 2.

4. DEFINING ILLOCUTIONARY ACTS

As we have seen, an illocutionary act is an institutional action performed by an utterance act (sending a message, in the case of artificial agents). We now face the problem of defining specific types of illocutionary acts, like assertions, promises, and requests. In view of our goals, we shall limit our treatment to illocutionary acts performed by artificial agents in an ACL.

To start with a simple example, let us consider an assertion. As a first attempt, we might stipulate that agent a's act of sending to agent b a message of the form

(assert a b s),

where s is a sentence in a suitable content language, counts as an illocutionary act of assertion, performed by a, addressed to b, and with a propositional content whose logical representation can be computed from s. Now the problem is: is this a correct way of defining assertions? Or should we look for a definition of asserting that is independent of the message level, and later define a "counts as" relation between messages and assertions? In search for an answer to this question, let us look to a well-known proposed standard.

4.1 An example from FIPA ACL

In FIPA ACL [9], an act of informing (an assertive illocutionary act) is defined as follows. First, informing is defined, independently of the message level, in terms of a set of feasibility preconditions (FP) and of a rational effect (RE). More precisely, if $B_a p$ means that agent a believes that p, we have:

FP $B_a p, \neg B_a(B_b p \vee B_b \neg p)$
RE $B_b p$.

The above FP and RE define an event type that, to keep to our notation, we shall here denote by $inform(a, b, p)$. The second step in the definition of informing, in FIPA ACL

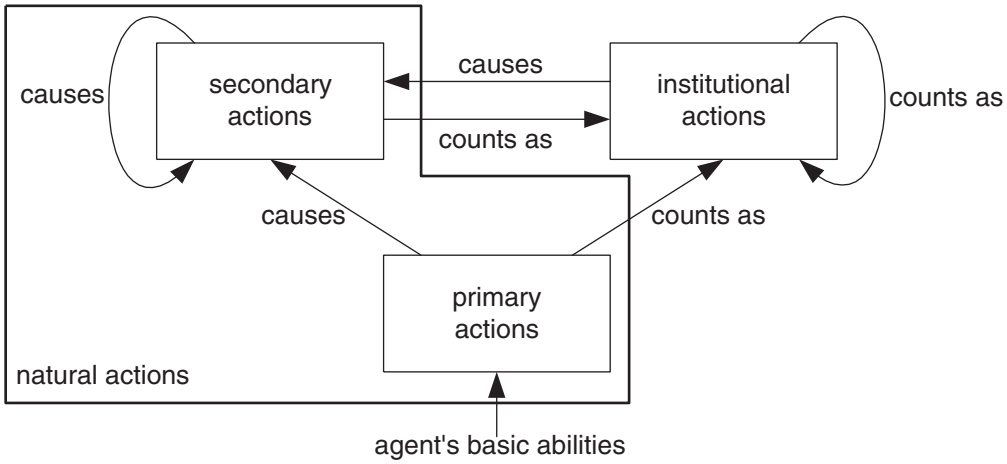

Figure 2: Types of actions and their realization relationships.

specification, is to define informing at the message level. What is needed here is a suitable syntax, including an illocutionary force indicator of informing. In FIPA's string encoding, an informing message has the form

(inform :sender a :receiver b :content s)

In our terminology, the relationship between sending such a message and informing can be described as follows: if p is a logical representation of the propositional content conveyed by sentence s, then by sending a message of the form given above agent a performs a primary action that counts as an illocutionary act of informing b that p.

FIPA's definitions have the merit of neatly separating the level of utterance (sending a message of a given form) from the level of illocution. In our opinion, however, they fail to recognize the institutional nature of illocution. Indeed, what does it mean for a message-sending event to count as an illocutionary act, if such an act is defined in terms of feasibility preconditions and of a rational effect? How can an event possibly "count as" a set of beliefs? We think it cannot, because a belief is not an institutional state.

The argument above can be extended to all illocutionary acts whose definition includes a reference to the mental states of the agents involved in a communicative exchange. If our characterization of illocution as institutional action is correct, it follows that mental states are not fit, or at least not sufficient, to define illocutionary acts. We believe that what is lacking in mentally oriented definitions of illocution is the social dimension of communication. This point will be clarified in the next subsection.

4.2 The social dimension of illocution

We now face the following problem: we want to deal with illocutionary acts so that their definitions are independent of the message level, but do not rely on mental states. We believe that the so-called social approach, already discussed in several scientific papers [19, 4], is the most promising direction, and we shall try to explain why.

From the perspective taken in this paper, an action is an event intentionally brought about by an agent, and an event is a change in the state of the world. So, in order to provide an analytical definition of illocutionary acts, we must first identify the kind of state change that is characteristic of illocution. The main tradition in AI models of illocutionary acts concentrates on changes at the level of agent mental states. Mental states are considered as essential because an agent that performs a speech act typically intends to affect the mental states of the addressee. But this does not mean that such changes should be used to *define* an illocutionary act. To understand this point better, consider again an example from the realm of natural actions. The typical aim for pressing a computer's ON/OFF button is to switch the computer on; but the act of pressing the button is defined independently of this aim. We think we should do the same with illocutionary acts; that is, we need definitions that are independent of the reasons why such acts are typically performed.

We believe that it is indeed possible to define illocutionary acts analytically in terms of state changes. The reason why this fact has been widely overlooked so far in AI is that the kind of states that have to be taken into account is foreign to the main AI tradition in agent modelling, even if it is rapidly gaining favor in the Multiagent System community [2, 5]. That is, illocutionary acts have to be defined in terms of changes at the level of the social relationship between agents. More precisely, illocutionary acts affect the network of *social commitments* that bind an agent to other agents. This point of view has already been defended by several authors [19, 4]. However, it has not yet developed in a full-grown theory of illocution. In the rest of this section we shall try to suggest how this could be done.

We take commitment to be a primitive concept underlying the social structure of multiagent systems. More precisely, a commitment is a social state that binds an agent (the *debtor*), relative to another agent (the *creditor*), to the fact that some proposition holds (the *content*). Let us consider again the example of assertions. We stipulate that $C_{ab}p$ means that agent a is committed, relative to agent b, to the fact that p holds. Then we define:

$$Event(e, assert(a, b, p)) \triangleq e]C_{ab}p.$$

The idea is that an event e is an assertion by a to b that p is the case, if and only if it commits a, relative to b, to the fact

that p holds. This definition is strictly analytical, in that it says that asserting is just making the specified commitment. But the definition also has a normative import, as we shall see in a while.

We now show how an assertion can be carried out by sending a message. To this purpose we stipulate that:

- a message is a piece of text sent by an agent, the *sender*, to another agent, the *receiver*;
- from a message it is possible to compute a logical representation of a proposition (the *content*) and an *illocutionary force indicator*;
- "assert" is an illocutionary force indicator.

If m is a message, we write $Event(e, send(a, b, m))$ to say that e is an event of agent a sending m to agent b. We also denote the message content as $cont(m)$, and the illocutionary force indicator as $illoc(m)$. We now state the following axiom:

$$CountsAs(e, send(a, b, m), assert(a, b, p)) \leftrightarrow$$
$$illoc(m) = "assert" \wedge cont(m) = p \wedge \Phi,$$

where Φ represents additional contextual conditions that a relevant institution may state for an assertive message to count as an assertion. (For example, it may be stated that both the sender and the receiver must be officially registered agents, that they must have previously opened a conversation, and so on.) These conditions may be dealing with roles, that is patterns of behavior agents must follow in order to respect the dictates of (electronic) institutions [6]. By applying Axiom A2, we are able to derive that sending a message, under given conditions, counts as an illocutionary act of asserting. From the analytical definition of asserting we then derive that the sender of the message has made a commitment relative to the receiver.

Now suppose that an agent makes two inconsistent assertions, so that we have:

$$Event(e_1, assert(a, b, p)) \wedge Event(e_2, assert(a, b, \neg p)).$$

Of course, p and $\neg p$ cannot hold at the same time. Should we conclude that a failed to assert that p? We believe not. What we conclude is that a has made two commitments that cannot be jointly *fulfilled*. Therefore, we know that at least one of the two commitments is *violated*. Fulfillments and violations of commitments are the basis for a normative treatment of communicative behavior; for example, we may take them into account to sanction undesirable behavior or at least to update an agent's reputation record.

The above definition of asserting is very simple, and may prove insufficient for practical applications. For example, it may turn out that some sort of "sincerity condition" (i.e., the fact that the sender actually believes what it asserts) is necessary to account for assertions. In such a case, we may define assertions as:

$$Event(e, assert(a, b, p)) \triangleq e](C_{ab}p \wedge C_{ab}B_a p).$$

It is important to stress that there is a sharp difference between this definition of sincerity and FIPA-like preconditions. With the definition above, an assertion can be successfully made even if the sender is insincere, because what is derived from it is not that the sender *is* sincere, but that it *commits*, relative to the receiver, to being sincere. An insincere assertion would still be analytically successful, even if normatively defective.

5. CONCLUSIONS

In this paper we pointed out the main differences between natural and institutional actions, and analyzed speech acts as a kind of institutional actions. More precisely, we have suggested that illocutionary acts should be modelled in terms of the social commitments brought about by their performance.

In our treatment of natural action we try to stress the causal role of intention, which is almost commonplace in the philosophical tradition but has not yet found its way in AI theories of action. We are aware that further progress in this line requires the development of a full-fledged theory of causation, which is by itself a very demanding task. However, we believe that any theory of action that leaves causation out is deemed to be incomplete, and incapable of distinguishing between natural and institutional actions.

Our approach to institutional action contrasts with the treatment suggested by Jones and Sergot [10], who propose to represent the "counts as" relation as a modal conditional operator. The main difference between the two approaches is that, contrary to the implications of our Axiom A2, Jones and Sergot's operator is non-monotonic. We agree with these authors that an irreducible non-monotonic component underlies human institutional action; for example, an apparently valid case of property transfer may turn out to be invalid if some further fact is taken into account. However, this need not be the case in a society of artificial agents. Moreover, if it turns out that the definition of some institutional action has to be non-monotonic, it is still possible to adopt our definition and to define the truth conditions of the formula $CountsAs(e, t, t')$ in some non-monotonic logic.

Once it is recognized that illocutionary acts are institutional actions, it becomes clear that they have to be defined in terms of their effects on some kind of institutional states. Social commitments come here as a very natural choice. However, even if several researchers have already put forward interesting proposals in this direction, much work has still to be done before a complete definition of a commitment-based ACL is available.

6. REFERENCES

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

[2] C. Castelfranchi. Commitments: from individual intentions to groups and organizations. In V. Lesser, editor, *Proceedings of the First International Conference on Multi–Agent Systems*, pages 41–48, San Francisco, CA, 1995. AAAI-Press.

[3] P. R. Cohen and C. R. Perrault. Elements of plan-based theory of speech acts. *Cognitive Science*, 3(3):177–212, 1979.

[4] M. Colombetti. A commitment–based approach to agent speech acts and conversations. In *Proc. Workshop on Agent Languages and Communication Policies, 4th International Conference on Autonomous Agents (Agents 2000)*, pages 21–29, Barcelona, Spain, 2000.

[5] B. Dunin-Keplicz and R. Verbrugge. Collective commitments. In M. Tokora, editor, *Proceedings of the Second International Conference on Multi–Agent Systems*, pages 56–63, San Francisco, CA, 1996. AAAI-Press.

[6] M. Esteva, J. A. Rodrguez, C. Sierra, P. Garcia, and J. L. Arcos. On the formal specifications of electronic institutions. In F. Dignum and C. Sierra, editors, *Agent-mediated Electronic commerce (The European AgentLink Perspective) LNAI*, volume 1991, pages 126–147. Springer-Verlag, 2001.

[7] R. E. Fikes and N. J. Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving. In *Advance Papers of the Second International Joint Conference on Artificial Intelligence*, pages 608–620, Edinburgh, Scotland, 1971.

[8] T. Finin, Y. Labrou, and J. Mayfield. KQML as an agent communication language. In J. Bradshaw, editor, *Software Agents*, pages 265–284. MIT Press, Cambridge, MA, 1995.

[9] FIPA. FIPA communicative act library specification. Specification, Foundation for Intelligent Physical Agents, www.fipa.org, 2000.

[10] A. Jones and M. J. Sergot. A formal characterisation of institutionalised power. *Journal of the IGPL*, 4(3):429–445, 1996.

[11] J. McCarthy and P. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In D. Michie and B. Meltzer, editors, *Machine Intelligence*, volume 4, pages 463–502. Edinburgh University Press, Edinburgh, MA, 1969.

[12] N. J. Nilsson. *Artificial Intelligence: A new synthesis*. Morgan Kaufmann Publishers, San Francisco, California, 1998.

[13] S. Russell and P. Norvig. *Artificial Intelligence: A modern approach*. Prentice Hall, Upper Saddle River, New Jersey, 1995.

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

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

[16] J. R. Searle. *Intentionality*. Cambridge University Press, Cambridge, UK, 1983.

[17] J. R. Searle. *The construction of social reality*. Free Press, New York, 1995.

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

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