A language for artificial agents

Studies in Communication Sciences 1 (2001) 1-32

MARCO COLOMBETTI*

A LANGUAGE FOR ARTIFICIAL AGENTS

Communication among artificial agents is a new subject of research that situates itself at the intersection of computer science with linguistics, philosophy, and logic. In this paper, I first introduce artificial agents, show why communication is a key aspect of their activity, and justify the use of speech acts as the basic unit of analysis of agent conversations. After a concise sketch of existing agent communication languages, I describe the difference between the two leading approaches to the definition of agent speech acts, namely the mentalistic and the social approach. In the second part of the paper, I distinguish various aspects related to the treatment of speech acts (ontological, semantic, normative, and practical) and show how the semantics of agent messages could be based on a suitable concept of commitment.

Keywords: agent communication language, speech act, commitment

Introduction

This paper is about agent communication languages, a new, intriguing subject of research that situates itself at the intersection of computer science, linguistics, philosophy, and logic. While the term "agent communication language" is fairly recent, the roots of this area of investigation can be traced back to the

* M. Colombetti, Faculty of Communication Sciences, Università della Svizzera italiana, Lugano (Switzerland), marco.colombetti@lu.unisi.ch; Politecnico di Milano, Milano (Italy), marco.colombetti@polimi.it

work on automatic understanding and generation of speech acts carried out in the field of artificial intelligence since the late seventies (Cohen & Perrault, 1979; Allen, 1983; Appelt, 1985; Airenti, Bara & Colombetti, 1993).

Up to now, research on agent communication languages has been done mainly by computer scientists, such as the author of this article. However, the kind of problems that the topic raises calls for a cooperation with linguists, language philosophers, and communication scientists in general. In particular, this paper originated from a seminar given by the author at the Istituto linguistico-semiotico (Linguistic-Semiotic Institute) of the Università della Svizzera italiana in Lugano, and it was written in hope to stimulate further discussions among scientists belonging to the different disciplines concerned with communication.

1. Agents and Communication

At the turn of the millennium, the World Wide Web has become part of everyday reality for a high number of human beings. Even its most enthusiastic supporter, however, will have to admit that the Web is a problematic object, because the huge amount of information it contains makes the use of it by human beings increasingly difficult.

Let us suppose, for example, that you belong to the recent, but already numerous, population of electronic auction addicts. You spend your nights browsing the Web because you hate the idea that you might fail to notice the bargain of your life. After your doctor has advised you to go back to a healthier lifestyle, you start wondering whether a software tool can browse the Web on your behalf to identify possible bargains, negotiate convenient prices, and finalize commercial transactions. That is, you want a system able to carry out a well-defined task in an autonomous way. To use a by now well-established term, you want a software *agent*.

The concept of an agent has been one of the most innovative notions to appear in computer science during the last decade. So innovative, I dare say, that nobody still understands it completely. This should come as no surprise. In fact, the term "agent" expresses the intuition that we would like software systems to share some important features with human beings: we would like them to be able to act rationally in a partially unknown and unpredictable world, populated by other agents that pursue their own, individual goals. This view is very different from the traditional perspective computer scientists have been raised to deal with.

Agents are interesting mainly because they interact with other agents. To carry out an electronic auction, for example, we need at least an agent in the role of auctioneer and a number of agents in the role of bidders. Every agent will strive to maximize its individual utility function, and therefore will have to be endowed with some degree of economic rationality in order to carry out its task in a satisfactory way. But this is not enough. A large part of the agents' interaction will consist of communication: for example, the auctioneer has to tell the bidders which items are put under the hammer; the bidders have to place their bids; the auctioneer has to assign an item to a bidder; the winner has to finalize the transaction; and so on. Therefore, we need to define languages for agent communication.

The task of designing an agent communication language may seem easy. After all, programs already communicate with each other, within a single computer, on local networks or via Internet. However, the situation with agents is different. We want them to be able to communicate with each other, and therefore we need to define one standard language, or at least a standard language for each community of agents. On the other hand, we do not want to severely limit *a priori* the spectrum of possible interactions, and we therefore need an open language, able to support a rich variety of communicative exchanges. This is the problem we shall deal with in the rest of the paper. We shall

start, in Section 2, by asking ourselves what should be the right unit of agent communication, and we shall see that there are good reasons to base an agent communication language on the notion of speech act. In Section 3, we shall have a quick look at some agent languages that have already been proposed and even adopted as standards *de facto*. In Section 4 I compare two approaches to the definition of semantics for an agent communication language. In Sections 5 and 6 I shall sketch an attempt to center the definition of speech acts on a suitable notion of commitment. Finally, in Section 7 I shall draw some conclusions.

2. The Unit of Communication

The observable, public component of a communication process is what we call a *dialogue* or *conversation*. As it happens with humans, a dialogue among artificial agents will have to be organized in *turns*, and each turn will consist of a sequence of units of some sort. The problem we now face is, What is the right sort of unit?

Before we try to answer this question, we must be aware that a communication process can be analyzed at different levels of abstraction. At a rather low level, for example, communication consists of sequences of characters transmitted through a physical connection. At a slightly higher level, communication consists of sequences of messages exchanged by a number of agents, a message being a well-formed sentence of some formal language. So far, nothing is specific to the idea of an agent communication language: messages are exchanged by all kinds of digital devices without any need to appeal to such a concept. We need, however, to move to a higher level of analysis. Why? Basically, because only a sufficiently abstract treatment allows us to assign suitable semantics to communicative units.

Let me clarify this point with an example. Suppose a digital thermometer is connected to a control unit, in charge of swit-

ching an air conditioner on and off. Every five minutes, the thermometer sends a message to the control unit, and communicates to it the current environmental temperature. In turn, when the temperature is too high (respectively, low) the control unit sends a message to the air conditioner, in order to switch it on (respectively, off). If we like, we can interpret all messages exchanged by the thermometer, the control unit and the air conditioner as speech acts: the thermometer *informs* the control unit that the temperature is high or low, and the control unit *requests* the air conditioner to switch on or off. However, such an interpretation adds nothing interesting. The whole system can be completely understood in purely causal terms: the message from the thermometer directly causes a predefined action of the control unit, and in turn the message from the control unit directly causes a predefined action of the air conditioner. Here, the notion of speech act is redundant.

With software agents, the situation is different. For example, consider an interaction between trading agents. An agent in the role of a buyer cannot directly cause the seller of a good to sell at a given price, because the seller itself is an autonomous agent. In other words, the seller will decide whether it will or will not sell according to an individual strategy that, in general, will be at least partially unknown to the buyer. This means that we cannot define the meaning of agent messages in terms of direct causal effects: we need to work at a higher level of abstraction.

The problem, now, is to identify the right level of abstraction for the treatment of agent messages. To do so, let us go back to the previous example. What kind of communicative acts do we expect a buyer and a seller to perform? We see that:

- the seller may *propose* to sell some goods;
- the buyer may *offer* to pay a certain amount of money for the goods on sale;
- the seller may *reject* or *accept* the buyer's offer;

and so on.

It is not difficult to identify the abstraction level at which we are describing the communicative interaction: it is the level of speech acts, more precisely of *illocutionary acts* (Austin, 1962; Searle, 1969). We therefore formulate the following working hypothesis: agent communication should be dealt with at the level of illocution.

As the reader will immediately see, we do not have a worked-out solution yet, but rather the statement of a problem. Several questions need to be answered, and among these:

- how should messages exchanged by software agents represent illocutionary force and propositional content?
- what repertoire of illocutionary acts is suitable for agent dialogues?
- how should software agents be internally structured to be capable of performing illocutionary acts?
- how can we define the semantics of agent messages in implementing illocutionary acts?

In the next section, I shall sketch some answers to these questions that have been given up to now.

3. A Glance to Existing Agent Communication Languages

The first agent communication language based on illocutionary acts has been KQML (Knowledge Query and Manipulation Language; Finin, Labrou & Mayfield, 1995), developed since the beginning of the nineties within the Knowledge Sharing Effort, a vast research program funded by DARPA (the US Defense Advanced Research Projects Agency). KQML messages implement a performative representation of illocutionary acts. For example, the KQML message,

(3.1) (tell
:sender *a*
:receiver *b*
:content *ProductionYear*(*o*,1792)
),

can be used by agent *a* to inform agent *b* that the production year of object *o* is 1792. More generally, the syntax of KQML specifies that a message starts with a *performative* (like `tell`), followed by a number of *parameters* (like `:sender`, `:receiver`, and `:content`), each parameter being in turn followed by a *value* (like *a*, *b*, and *ProductionYear*(*o*,1792)).[1] Two points are of particular interest: the performative form of messages, and the choice of the *content language*, that is, the formal language used to represent the content of a message.

Let us analyze the first point, that is, the representation of illocutionary force by an illocutionary verb in performative form. This choice is motivated by the desire to avoid ambiguity in the representation of illocutionary force. For example, in ordinary language an utterance in the future tense (like "I shall be in Lugano tomorrow") can be used to express an expectation or to make a promise, and only the context of the utterance allows the addressee to understand the illocutionary force intended by the speaker. It is well known, in particular to those who work in the field of automatic language processing, that the use of context to disambiguate illocutionary force is a difficult task. The use of explicit performatives is intended to eliminate such a difficulty.

Once it is decided that illocutionary forces are represented in this way, it is necessary to choose a set of allowable performatives. In KQLM, there are 43 performatives, divided in 11 catego-

[1] The syntax adopted by KQML to express the content of a message is somewhat different from the one I use in this paper. The difference, however, is irrelevant for our current purposes.

ries: basic informative (3 verbs), database (4), basic response (2), basic query (7), multi-response (3), basic effector (2), generator (6), capability-definition (1), notification (2), networking (7), and facilitation (6). Here are some examples (see KQML's specification at `http://www.cs.umbc.edu/kqml/`):

- basic informative performatives: `tell`, `deny`, `untell`;
- basic query performatives: evaluate, reply, ask-if, ask-about, ask-one, ask-all, sorry.

As can be seen even from this short list, the categories of performatives do not correspond to classical taxonomies of speech acts based on illocutionary force, but are grouped according to contexts of use. For example, `reply` is used to perform an assertive act and `ask-if` to perform a directive, but they are grouped together because they both pertain to the process of questioning and answering.

The values of parameters of KQML messages are represented according to a suitable formal syntax. In particular, contents are represented by formal sentences in KIF (Knowledge Interchange Format), an extension of a first-order predicate language developed within the already mentioned Knowledge Sharing Effort. For simplicity's sake, in this paper I shall adopt a more classical logical notation.

Since its definition, KQML has become a standard *de facto* in the development of agent systems. More recently, the Foundation for Intelligent Physical Agents (FIPA) has proposed a new standard, named ACL (FIPA, 1997). For the aims of this paper there is no important conceptual difference between KQML and ACL, and therefore I shall not describe ACL here.

So far I have said nothing about the semantics of a language like KQML. First, let me stress that an agent communication language should have well-defined formal semantics. The reason is that software agents must be designed to produce and analyze messages correctly, and this would be almost impossible without an unambiguous, formal definition of the semantics of the communication language. Such a definition is not a problem

as far as the content language is concerned. As I have already remarked, the content language of KQML, KIF, is a first-order predicate language, and its semantics can therefore be defined following the methods of symbolic logic. But how can we account formally for the illocutionary force of messages? In other words, how do we define the semantics of performatives?

So far, two different approaches that have been suggested. The first approach, which I shall call *mentalistic*, assumes that semantics can be defined in terms of agents' *mental states*, like beliefs, desires and intentions. The second approach, which I shall call *social*, assumes that semantics requires a definition of the *commitments* brought about by the performance of a speech act.

Both KQML and ACL have been given mentalistic semantics (see Cohen & Levesque, 1995, and Labrou & Finin, 1997, for KQML; FIPA, 1997, for ACL). From an ontological point of view, this means that agents are assumed to be the kind of entity that can entertain mental states – a standpoint that is coherent with the mainstream artificial intelligence treatment of agents as artificial reasoners (see for example Wooldridge, 2000).

From a technical point of view, mental states (also known as *propositional attitudes* in analytic philosophy) can be represented by sentences of a first-order modal logic (see for example Hughes & Cresswell, 1996). For example, the fact that agent a believes that the production year of object o is 1792 can be represented symbolically by:

$$B_a ProductionYear(o,1792),$$

where B_a is the so-called *epistemic operator*, expressing belief: if a is an agent, and φ is a formal sentence representing a proposition, the modal sentence $B_a\varphi$ says that a believes φ.

Appropriate axiom systems are used to capture the essential properties of mental states. For example, the *axiom of coherence*,

$$B_a\varphi \rightarrow \neg B_a \neg \varphi$$

(read: if a believes φ, then a does not believe not-φ), specifies that a rational agent should always entertain consistent beliefs. Such an approach is not without problems; in particular, we are far from having a satisfactory axiomatic treatment of volitional mental states, like desire and intention. But, for the moment, let us leave these problems aside, and see how we can define illocutionary acts in terms of mental states. To do so, let us take a simple example, the illocutionary act of *informing*. Following the mentalistic approach, we can define "a informs b that φ" as the act that has the following *preconditions*:

(3.2) $B_a\varphi$ (a believes φ),
(3.3) $\neg B_a B_b \varphi$ (a doesn't believe that b believes φ),

and the following *expected effect*:

(3.4) $B_b\varphi$ (b believes φ)

(see for example FIPA, 1997). Sentence 3.2 corresponds to the sincerity condition of an assertive act in Searle's speech act theory. Sentence 3.3 corresponds to one of Searle's preparatory conditions (see Searle & Vanderveken, 1985). Finally, sentence 3.4 captures the *perlocutionary effect* that a speaker typically intends to achieve by informing an addressee about some fact.

Although informing is just one among many types of illocutionary acts we may want to define, its formal definition already shows both the strength and the weakness of mentalistic definitions. The strength is that the definition of illocutionary acts needs no *ad hoc* apparatus: in artificial intelligence, mental states are used to describe rational agents in general, and are not introduced especially for modeling communication.

There are, however, several weak points. The first is that the status of conditions like 3.2 and 3.3 is not clear. What does

sentence 3.2 mean, after all? Does it imply that agents are sincere (by constitution, so to speak)? Or that they ought to be sincere? Or that we expect them to be sincere? The second weak point is even more problematic. If we limit our ontology to mental states, the effect of an illocutionary act can only be defined in terms of expected perlocution: informing amounts to an attempt to convince. Now, what happens if the expected effect is not achieved, that is, if *a* does not convince *b* about φ? Should we say that *a* did not inform *b* that φ? This does not sound correct. After all, believing or not what *a* says is part of *b*'s private business. But then, if sentence 3.4 is not essential to define an act of informing, why should it be part of its formal definition?

According to some authors (Singh, 1998; Colombetti 1999, 2000), these problems show that there are difficulties for a strictly mentalistic approach: by themselves, mental states are intrinsically insufficient to define illocutionary acts. To go back to the previous example, a component, which the mentalistic definition of informing completely lacks, is the sender's commitment to what it says. By this I mean that when an agent, *a*, informs another agent, *b*, that φ is the case, then *a*'s act creates some form of obligation for *a* with respect to *b*. Such a notion cannot be found in the currently existing semantics of either KQML or FIPA ACL.

There are, in fact, formal languages that include an agent communication component whose semantics are defined in terms of some notion of commitment. Examples of these are Elephant 2000[2] (McCarthy, 1990) or Agent-0 (Shoham, 1993). In my opinion, however, the authors of these languages fail to do justice to the conceptual complexity of commitment, a notion

[2] Elephant 2000 is so named because it is proposed as a prototype language for the next Millennium and because "it never forgets anything". This means that the commitments created by speech acts are automatically stored in memory and never expire.

that brings in a bundle of problems connected with the normative aspect of social interactions.

More recently, several researchers have stressed that agents are social entities, involved in social interactions that include communicative exchanges (see for example Conte & Castelfranchi, 1995). Some authors have also started to propose agent communication languages based on articulated treatments of social interactions (Singh, 1998). However, no such proposal has yet gained universal approval.

In the next section, I shall compare the mentalistic and the social approaches to agent communication, in order to clarify their pros and cons.

4. Mental and Social Aspects of Communication

If we look back at speech act theory, as it has been developed by philosophers and linguists, it appears that the definition of the illocutionary acts performed in natural languages involves both mentalistic and deontic concepts. However, I think we should resist the temptation to immediately extend this standpoint to agent communication languages. Agent communication is going to be much simpler than human communication, and therefore a different approach to the definition of illocutionary acts might be preferable. In particular, I think we should try to avoid, or at least to limit severely, the mentalistic component in the definition of agent speech acts. There are several reasons to do so: Different agents might have completely different internal structures, and this is going to make the definition of a standard set of mental states extremely difficult. For example, all agents are likely to have some internal state that represents information about the environment, and this we can reasonably call "belief". But this is not necessarily the case with all types of mental states. For example, some agents might distinguish between desires, goals and intentions, and other agents might not do so.

Even if we can choose a standard set of basic mental states, we do not know yet how to define them formally. With the exception of beliefs, for which there are treatments accepted by the majority of researchers, the formal theories of mental states are still controversial. The mental states of an agent are typically unobservable by other agents, and this is going to bring in severe difficulties as far as understanding and reacting to speech acts is concerned. For example, under what conditions is an agent going to assume that another agent really believes what it says?

The social approach, based on the notion of commitment, does not suffer from all these difficulties. Because deontic states are external and public, the social approach is insensitive to differences in the internal structure of agents, and avoids the difficulties deriving from the unobservability of mental states. In principle, the description of all agents' commitments can be kept in a public store, and accessed by every agent at any moment. An agent cannot know whether another agent is sincere, but it certainly can know whether the other agent has made a specific commitment.

Let me now insist on the difference between the two approaches using message 3.1 as an example. According to the mentalistic approach, agent *a* will send message 3.1 when it believes that the production year of object *o* is 1792, it believes that agent *b* does not yet believe that the production year of *o* is 1792, and it wants *b* to believe that the production year of *o* is 1792. After receiving the message, agent *b* will typically assume that all such conditions hold. In particular, *b* will now believe that *a* believes that the production year of *o* is 1792. If *b* assumes that *a* has access to the right source of information on the production years of objects, *b* will come to believe that the production year of *o* is indeed 1792, and the goal of *a*'s speech act will be achieved. This picture is, however, too idyllic. In particular, it does not take into account the possibility that *a* lies about the production year of *o* — a possibility which cannot be ruled out in a competitive situation like a commercial transaction.

Let us now see how the same message could be dealt with without relying on mental states. First, we assume that agents interact in a context in which it is meaningful to say that an agent has a commitment with respect to another agent. After sending message 3.1 to agent *b*, agent *a* will be committed, relative to *b*, to the fact that the production year of *o* is 1792. That's it: for the moment, we do not need to say anything else. We do not consider, at this level of analysis, why agent *a* intended to take up such a commitment, nor if and how agent *b* will react to it. These aspects are important to understand the interaction between *a* and *b*, but are not part of the semantics of message 3.1.

The social approach is attractively simple. However, it brings in a number of difficult questions, and in particular:

- how can we create contexts in which it is meaningful for an artificial agent to make a commitment relative to another agent?
- what should happen if a commitment is not fulfilled?
- can we define all relevant types of illocutionary acts without relying on mental states?

In the next two sections, I shall sketch a possible approach that looks promising to me.

5. Communication as a Social Activity

When we say that communication is a social activity, we do not only mean that communication is a process that takes place in a group of agents. We also want to stress that communication involves a number of social institutions, which specify and regulate the commitments created by communicative acts.

From now on, I shall assume that the characteristic function of communication is to create (or cancel, or modify) commitments, involving the sender and the receiver(s) of a message (and possibly third parties, that do not participate to the commu-

nicative exchange but are referred to in the message or somehow belong to its context). If this assumption is correct, to account for communication in a satisfactory way we have to specify:

(i) what kinds of commitments can be made;
(ii) how a speech act can create a specific commitment;
(iii) what consequences the creation of a commitment has for an agent;
(iv) how these consequences can be taken into account by an agent to act rationally.

Point (i) has to do with the *ontology* of communication, in that it clarifies what kind of social facts are presupposed by communication. Point (ii) regards the *semantics* of messages, in that it specifies how a message can create a particular commitment. Point (iii) has to do with the *normative systems* regulating commitments. Finally, point (iv) concerns what I call the *practical aspect* of communication (Colombetti, 1999), that is, the connection between communication and rational action. In the rest of this section I shall separately analyze these four points.

5.1. Commitments

When I say that commitment is part of the ontology of speech acts, I mean that commitments must logically pre-exist if we want speech acts to be possible. For example, asserting something involves a commitment to the truth of what is asserted. Without the possibility of committing to the truth of a statement, there could be no assertions.

From a logical point of view, commitment can be regarded as a deontic concept — that is, as a concept somehow related to obligation. However, present day deontic logic (i.e., the branch of modal logic that deals with obligations, permissions, and so on) does not offer a formal treatment of commitment. In the following, however, I shall suggest possible ways of dealing with commitment using the methods of classical modal logic.

In deontic logic (see for example Åqvist, 1984), it is shown that all basic deontic states can be reduced to two fundamental notions, which we can name "necessity" and "violation" (Anderson, 1958). A state, described by sentence φ, is *obligatory* if and only if $\neg\varphi$ (i.e., the negation of φ), necessarily implies a violation. We shall now write this definition in symbolic form. We use the modal operator O to mean that something is obligatory, the modal operator $\Box$ to mean that something is necessary, and the propositional symbol V to denote a violation. Taking $\Box$ and V as primitive notions, we can define "φ is obligatory" as:

$$O\varphi =_{\text{def}} \Box (\neg\varphi \rightarrow V).$$

For example, let us consider the sentence "it is obligatory to pay taxes". If we represent "to pay taxes" by the symbol *PayTaxes*, we can express this obligation by

$$O\ PayTaxes.$$

Our definition of O tells us that this statement is equivalent to

$$\Box (\neg PayTaxes \rightarrow V),$$

which means that not paying taxes necessarily implies a violation.

At first sight, we might think that a violation coincides with liability to a sanction. Indeed, there is a strict practical relationship between violations and sanctions. Most violations lead to some kind of sanction, and therefore it is important to see how sanctions can be concretely associated to violations (see Subsection 5.2). Moreover, as we shall see in Subsection 5.4, avoiding the sanctions involved by violations may be a major reason for a rational agent to follow regulations. However, at least in

principle, the concept of violation appears to be ontologically prior to that of sanction.[3]

The question now is, Can we define commitment along similar lines? I think we can. First, let us consider the intuitive differences between obligations and commitments. In general, an obligation derives from some general law, which applies to all subjects that meet certain conditions. For example, male Swiss citizens are obliged to serve in the Swiss Army according to certain regulations, just because they are male Swiss citizens. On the contrary, commitments are deontic states that do not derive from general laws, but are typically created by individual actions. Moreover, commitments apply to specific individuals and are relative to some other individual or group of individuals. For example, if I promise to my wife that I will make dinner tonight, this very act of promising brings about a commitment that binds me on this particular occasion and relative to my wife. We therefore conclude that a commitment is always a commitment *of* some agent *a*, *relative* to some agent *b*.[4] As in the case of obligation, we can define commitment in terms of violation. In this case, however, a violation will have to be a violation by some agent *a*, relative to some agent *b*. I shall therefore write

$$V_{ab}$$

---

[3] There can even be violations without sanctions. About twenty-five years ago, a traffic regulation was issued in Italy, that obliged all car owners to apply a sticker to their cars, showing the maximum speed allowed for that type of car. However, there was no sanction for not doing so. Needless to say, most car owners did not apply the sticker. They violated the traffic regulations, but were not liable to any sanction.

[4] In general, *a* and *b* will be distinct agents. As a special case, however, an agent can assume a commitment relative to itself. Self-commitments may require a special treatment, and I shall not deal with them in this paper.

to say that agent *a* has violated a commitment relative to agent *b*. From this, we can define an *indexed commitment operator*

$$C_{ab}\varphi =_{\text{def}} \Box (\neg\varphi \rightarrow V_{ab}),$$

where $C_{ab}\varphi$ means that agent *a* is committed to φ relative to agent *b*.

It is now time to go back to our main topic, that is, speech acts. The connection between speech acts and commitments is very simple: speech acts are the means by which commitments are brought about. In the next subsection we shall see how this can be done.

5.2. The semantics of messages

An agent performs an illocutionary act by executing another action at a lower level of abstraction – namely, by sending a message to another agent (this action roughly corresponds to Austin's locutionary act and to Searle's utterance act). In this paper, I define the *semantics* of a message to be the function that maps a message (and its context of performance) into the commitments brought about by sending the message. To make an example, semantics has to specify what commitment agent *a* brings about when it sends the message

(5.1)
```
(assert
    :sender a
    :receiver b
    :content ProductionYear(o,1792)
).
```

I shall say that by producing message 5.1, agent *a* enters a state such that the falsity of *ProductionYear*(*o*,1792) necessarily implies a violation by *a* relative to *b*. Following the lines of the previous subsection, we can then express such a state as:

C_{ab}*ProductionYear*(*o*,1792).

Now, what happens if the information provided turns out to be false? From our definition of obligation we derive a violation

V_{ab}.

More generally, we can define the semantics of all messages of this kind by saying that a message of the form

(5.2)
```
(assert
    :sender a
    :receiver b
    :content φ
)
```

brings about that

$C_{ab}\varphi$.

In my opinion, the above definition is very reasonable, and allows us to avoid the use of mental states. On the other hand, if we prefer to assume that all agents must be able to entertain beliefs, we can provide an alternative definition, by stating that sending message 5.2 has the effect that

$C_{ab}B_a\varphi$,

that is, that *a* is committed to believing that φ.

In Section 6 we shall see how this approach to the definition of the semantics of messages can be applied to illocutionary acts of different types.

5.3. Normative systems

In practical situations, it is not sufficient to know that a violation took place. One also needs to know what kind of sanction should be applied as a penalty. This is particularly evident in situations that involve legal effects, where violations are typically sanctioned by law: think for example of a false statement given in the context of a commercial transaction. On the other hand, there are many situations in which the sanction for a violation is not specified by a law, but still plays an essential role in regulating human interactions. If you lie to your spouse about what you did last night, and your spouse understands this, s/he is likely to apply some kind of sanction, even if no law explicitly forbids lying to spouses.

Dealing with artificial agents, I assume that violations are regulated by sets of norms, which I call *normative systems*. To clarify this point, let me give a concrete example. Suppose that we establish a normative system, which we call *comm-trans*, to regulate commercial transactions among agents. In particular, within this normative system we want to define specific sanctions for specific violations. For example, we want to establish that giving false information about the production year of an object is sanctioned by paying 100 euros to the agent that received the false piece of information.[5] To specify this in an unambiguous way, we might include a reference to a specific normative system, n, in the symbol denoting a violation, which now becomes

$$V_{abn}.$$

Such a reference allows an agent to access a knowledge base describing a specific normative system and specifying what

[5] In fact, the 100 euros will not be paid by the artificial agent itself, but by a human being, that is, by the *owner* of the agent.

sanction is to be applied for every specific violation. In the specific case of our example, this norm may be denoted by the formal term *comm-trans*(*prod-year*). A knowledge base,[6] accessible to all trading agents, will then specify that the sanction for the violation described by

$$V_{a,b,comm\text{-}trans(prod\text{-}year)}$$

is that agent *a* pays 100 euros to agent *b*.[7] Being aware of the sanctions associated to a violation may have an important impact on the behavior of agents, as I shall show in the next subsection.

5.4. Practical reason

In the field of artificial intelligence, agents are conceived as *rational systems*, that is, as systems that have goals and are able to plan their activity in order to reach as many goals as possible. Equivalently, we can regard agents as systems that have a utility function to maximize, and build and execute action plans that allow them to achieve sufficiently high values of such a function. By this, agent rationality is reduced to *economic rationality*, and an agent will typically execute an action plan when this leads to a higher reward than the execution of alternative plans (including the empty plan, which amounts to doing nothing). As a consequence, an agent will perform a speech act (typically, as part of a larger action plan) because it expects some reward from its execution.

---

[6] A *knowledge base* is a set of formal statements from which a software system can derive conclusions in a purely mechanical way.

[7] This example should not be taken too seriously. It is only meant to show how one may concretely specify penalties for violations. There might well be more effective ways of doing so.

It is in the context of this view that we can consider important issues like the sincerity of an assertion. Agents may not have any compulsion to sincerity. But, in general, we can expect an agent to be sincere when this leads to a higher reward than lying. For this reason, it is very important to define sanctions for insincere assertions: penalties for lying can be taken into account by rational agents, which may avoid lying just because it is too costly.

It is now time to go back to the limitations of mentalistic models of speech acts, already mentioned in Section 4. In a mentalistic model, the sincerity condition of an assertion is typically viewed as a condition deriving from principles of rational behavior. In my opinion, this statement is empty, unless we can describe the process that leads an artificial agent to be sincere in order to maximize its utility function. But all current mentalistic models of agent communication fail to do so. On the contrary, the model proposed in this paper explains sincerity as the attempt to avoid the sanctions associated to lying. Similar considerations apply to non-assertive speech acts, which I shall define in the following section.

6. Further Speech Acts

So far I have based my approach to agent communication languages on a single type of illocutionary act, namely assertion. In this section I deal with further assertive acts (like the acts of informing, confirming, and so on) and with non-assertive speech acts. Throughout the section, I shall adopt the classification of illocutionary acts proposed by Searle (1975).

6.1. More assertive acts

In FIPA ACL three basic information passing acts are defined, namely *informing*, *confirming*, and *disconfirming*. It is easy to

see that the difference among the three types of acts concerns only what, in Searle's terminology, can be viewed as a preparatory condition. Such conditions are:

- in the case of *informing*: that the receiver does not already know what is asserted by the sender;
- in the case of *confirming*: that the receiver already knows what is asserted by the sender, but may be uncertain about it;
- in the case of *disconfirming*: that the receiver believes that what is asserted by the sender is false.

In my opinion, there are two shortcomings in this approach. The first is that an agent might want to perform an assertive act without specifying whether it is an instance of informing, confirming or disconfirming. In other words, I think agent should be able to perform a more neutral kind of assertive act, which we can simply call `assert`.

In some cases, however, it may actually be relevant for an agent to make it explicit that it is asserting something in order to confirm or disconfirm a previous belief of the addressee. For example, agent *a* may want to tell agent *b* that the production year of object *o* is 1972, in order to disconfirm *b*'s previous belief that the production year of *o* is 1792. Now, it may well be that the difference between asserting and confirming (or disconfirming) cannot be defined without explicit mention of a mental state of belief. If this were true, the social approach alone would be intrinsically insufficient to define the illocutionary force of certain speech acts. But there is also a different approach, which seems to me more suitable for artificial agents. Suppose that, on a previous occasion, agent *a* has asserted to *b* that the production year of *o* is 1792. Later on, *a* discovers that the production year of *o* is indeed 1972, and may want to change its state of commitment relative to *b*. An agent communication language may provide a speech act for doing so. Such a speech act can be regarded as a case of disconfirming, and can been defined without taking *b*'s beliefs into account.

I do not know whether all cases can be treated in a similar way. In any case, whether agent communication can be completely dealt with in terms of commitments is an important topic for future research.

6.2. Commissives

The most common example of a commissive act is a promise. However, promising has a special condition, in that it presupposes that the promised act is advantageous for the receiver. A more neutral type of commissive act, that we may simply call *commit*, does not rely on this assumption (see Searle & Vanderveken, 1985). A commissive act made by a to b can be defined by two conditions:

- *propositional content condition*: the content is a statement of the form $Do(a,\alpha)$ describing an action of type α to be performed by a;
- *deontic effect*: $C_{ab}Do(a,\alpha)$; that is, a is committed, relative to b, to performing an action of type α.

As syntax, we can adopt any form that makes the commissive illocutionary force manifest. For example:

```
(commit
  :sender a
  :receiver b
  :content Do(a,α)
).
```

In most practical cases, a commissive will also include constraints as to when the action will take place. Such qualifications can be included in the description of the action (i.e., in the formal expression α), following some suitable syntax.

6.3. Directives

The most common example of a directive act is a request. Like in the case of commissives, however, requesting has a special condition, in that it leaves it open for the receiver to accept or reject the directive. A more neutral type of directive act, that we may simply call *direct*, does not rely on this assumption (see Searle & Vanderveken, 1985). The propositional content condition, quite obviously, is that the content of a request is a statement of the form $Do(b,\alpha)$, describing a future action of type α to be performed by b. A suitable syntax for directing could be:

```
(direct
  :sender a
  :receiver b
  :content Do(b,α)
).
```

The effect of a directive, however, is less obvious. So far, the effect of a speech act has always been defined as a commitment *of the sender*. Can we define directives along similar lines?

In fact, we would like a directive to imply a commitment *for the receiver* of the message. But how can a speech act performed by an agent create a commitment for another agent? In human communication, this would require a specific kind of relationship between the sender and the receiver, which has to be included in the definition of the speech act.

It seems to me that directives addressed by artificial agents to other artificial agents will have a feature in common with human directives: that is, they will be made within a predefined context of interaction, which will make it feasible for the sender to create an obligation for the receiver. For example, let us consider an agent, a, which intends to buy a copy of a book and another agent, b, which has that book for sale. We expect a to request b to sell the book, and b to accept a's request. The important point is that a's request is made within a predefined context

of interaction, defined by what we may call the *agreement* between booksellers and potential buyers. From the standpoint of this paper, an agreement specifies a set of *conditional commitments* of at least two agents. A typical agreement between a bookseller and a potential buyer may state that:

- the bookseller, *b*, is committed to delivering a book to a potential buyer, *a*, under the conditions that *b* has the book for sale and that *a* commits to paying for the book;
- in turn, *a*'s commitment to paying for the book is conditioned by the fact that *a* receives the book.

At the moment, this is not much more than a working hypothesis, but it seems to me a promising approach to the treatment of directives for artificial agents (more on this in Section 6.6).

6.4. Proposals

Offers and proposals are a common and important component of agent interactions. It seems to me that such speech acts can be defined as *conditional commissives*, that is, as implying the commitment for the sender to perform some future action, under the condition that the receiver assumes some other commitment. Let us consider an example: agent *a* may propose to buy an object, *o*, from agent *b* at the price of 5 euros by producing the message:

```
(propose
  :sender a
  :receiver b
  :content Do(a,buy(a,b,o,5,euro))
).
```

It is feasible to make such a proposal because buying is, by definition, an *interaction*, in which an agent transfers the property of some object to another agent in exchange of a sum of money. In other words, *buy*(*a*,*b*,*o*,5,*euro*) involves two distinct actions: *a*'s transfer of 5 euros to *b*, and *b*'s transfer of *o* to *a*.

Through its proposal, *a* commits to paying 5 euros on condition that *b* commits to transferring the property of *o* to *a*.

The logical analysis of conditional commitment might be carried out along the lines of treatment provided by deontic logic for conditional obligation (see again Åqvist, 1984). I shall not deal here with this technical aspect, which is beyond the scope of this paper.

6.5. What about expressives?

Expressive illocutionary acts are used by humans to express feelings and psychological states. Examples of expressives are, "Congratulations for winning the Nobel prize" and "I apologize for breaking your Ming vase". In human interaction, expressives appear to be an essential device to define and maintain interpersonal relationships. However, I do not see why expressive acts should be of interest for artificial agents. At least for the moment, we can completely neglect this category of illocutionary acts.

6.6. Declarations

Declarations are illocutionary acts that, by convention, create some institutional state of affairs. A typical declaration in a society of trading agents can be, "The auction is open", which may be implemented by a message like:

```
(declare
  :sender a
  :receiver b
  :content Open(auction)
).
```

The declaration must be produced by an agent with the required authority. The semantics of a declaration do not involve a commitment, but rather the creation of the relevant state of

affairs. For example, the semantics of the above declaration is that *Open*(*auction*) becomes true as an effect of its performance.

It is interesting to note that, in principle, all speech acts in performative form can be treated as declarations (see Searle & Vanderveken, 1985). This means that an agent communication language like the one I have sketched so far could be defined starting from only two primitives: the act of declaring and the operator of commitment. All other speech acts can be introduced in the language through internal definitions (what computer scientists usually call a *macro* definition). Here are the definitions of some of the acts I have introduced in the previous subsections (for conciseness' sake, I drop the keywords `:sender`, `:receiver` and `:content`):

(assert a b φ)
$=_{\text{def}}$ (declare a b $C_{ab}\varphi$)

(commit a b $Do(a,\alpha)$)
$=_{\text{def}}$ (declare a b $C_{ab}Do(a,\alpha)$)

(direct a b $Do(b,\alpha)$)
$=_{\text{def}}$ (declare a b $C_{ba}Do(b,\alpha)$)

It is not difficult to check that these definitions attribute to messages the same semantics as previously defined. Apart from its conceptual interest, the reduction of all non-declarative acts to declarations seems to be a powerful device for extending an agent communication language with new kinds of speech acts when required by applications.

Another advantage of a declaration-based approach is that it aids in the understanding of differences between commissives and directives. As I have already said, a declaration must be produced by an agent *with the required authority*. For example, if you meet your two friends Ann and Bob walking in the town park and you say "I pronounce you man and wife", there will be no institutional consequence. Analogously, agent a cannot create a commitment for agent b to do α by saying "I hereby commit

you to do α", unless a has the right authority to create a commitment for b. As suggested in Section 6.4, a may have such an authority on the basis of a predefined agreement.

7. Final Thoughts

In this paper I have tried to introduce the reader to the problem of agent communication, to compare the mentalistic and the social approaches to the definition of speech acts, and to outline how the social approach could be concretely carried out. In particular, I have tried to show how commitments can be used to define the semantics of agent messages, and to point out some aspects that urgently require theoretical and empirical work.

It is important to understand that even if I propose to define the semantics of agent messages in terms of commitments and without relying on mentalistic concepts, I do not intend to claim that all aspects of agent behavior can be understood in non-mentalistic terms. To give an example, it is obvious that to define *lying* we need to take beliefs into account: an agent can be said to lie about φ when it commits to the truth of φ and simultaneously believes that not-φ. However, lying is not itself a speech act, but rather an action that can be performed by means of an assertive speech act — and of course nobody would seriously think of defining a performative of the form "I hereby lie that...". So, my suggestion to avoid mentalistic concepts should be understood as limited to the semantics of messages.

When confronted to commitment-based semantics, many researchers feel that there might be problems to define commitment as a self-standing concept. Two objections are most common:

(i) that the very notion of commitment can be reduced to individual mental states, and
(ii) that an agent can commit to the performance of an action but cannot commit to the truth of a sentence.

Let me deal briefly with these objections. As regards (i), I think there is no way to eliminate the notion of commitment by defining it in terms of mental states that do not themselves involve some primitive deontic concept. For example, one could suggest that $C_{ab}\varphi$ actually means that both a and b believe that not-φ implies a violation. Such a definition, however, is based on violation and therefore does not eliminate the deontic dimension. As far as objection (ii) is concerned, I see no problem in the idea that an agent can commit to the truth of any sentence: this only means that if the sentence happens to be false, a violation is brought about. It is true, however, that the type of commitment involved in assertives appears to be different from the type of commitment involved in commissives. In speech act theory, the difference between assertives and commissives is basically one of *direction of fit* (Searle, 1969): while a commissive is fulfilled if the world is made to satisfy the content, an assertion is true if its content corresponds to the actual state of the world. I think that this concept can be extended to commitments. However, it is not yet clear whether this is necessary for a suitable formal treatment of agent speech acts.

Finally, let me remark that only real applications can show whether a purely social approach to agent communication is feasible. As far as we know today, it might well turn out that, as it happens with human languages, we need a combination of mental and deontic concepts to define the semantics of illocutionary acts performed by artificial agents. But much further work is needed before we can clarify our ideas on this matter. In any case, even if the last word will be to real applications, it seems to me that the definition of a suitable agent communication language cannot be left to computer scientists alone. The theoretical problems at stake urgently call for contributions by specialists of different disciplines concerned with communication.

References

AIRENTI, G., BARA B.G. AND COLOMBETTI M. (1993). Conversation and behavior games in the pragmatics of dialogue. *Cognitive Science* 17: 197–256.

ALLEN, J.F. (1983). Recognizing intentions from natural language utterances. In: M. BRADY AND R.C. BERWICK (eds.). Computational models of discourse, Cambridge, MA: MIT Press.

ANDERSON, A.R. (1958). A reduction of deontic logic to alethic modal logic. *Mind* 67: 100–103.

APPELT, D. (1985). Planning English sentences, Cambridge, UK: Cambridge University Press.

ÅQVIST, L. (1984). Deontic Logic. In: D. GABBAY AND F. GUENTHNER (eds.). Handbook of philosophical logic II, Reidel, pp. 605–714.

AUSTIN, J.L. (1962). How to do things with words, Oxford, UK: Clarendon Press.

COHEN, P.R. AND LEVESQUE H.J. (1995). Communicative actions for artificial agents. Proceedings of the International Conference on Multi-Agent Systems, San Francisco: AAAI Press.

COHEN, P.R. AND LEVESQUE H.J. (1990). Rational interaction as the basis for communication. In: P.R. COHEN, J. MORGAN AND M.E. POLLACK (eds.). Intentions in communication, Cambridge, MA: MIT Press, pp. 221–256.

COHEN, P.R. AND C.R. PERRAULT (1979). Elements of a plan-based theory of speech acts. *Cognitive Science* 3: 177–212.

COLOMBETTI, M. (1999). Semantic, normative and practical aspects of agent communication. Pre-prints of the IJCAI'99 Workshop on Agent Communication Languages, Stockholm, 51–62.

COLOMBETTI, M. (2000). A commitment-based approach to agent speech acts and conversations. Pre-prints of the Autonomous Agents 2000 Workshop on Agent Languages and Conversation Policies, Barcelona, 21-29

CONTE, R. AND C. CASTELFRANCHI (1995). Cognitive and social action, London: UCL Press.

FININ, T., LABROU Y. AND MAYFIELD J. (1995). KQML as an agent communication language. In: J. BRADSHAW (ed.). Sofware agents, Cambridge, MA: MIT Press.

LABROU, Y. AND FININ T. (1997). Semantics and conversations for an agent communication language. Proceedings of the Xth International Joint Conference on Artificial Intelligence (IJCAI'97), Nagoya, Japan.

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

HUGHES, G.E. AND CRESSWELL M.J. (1996). A new introduction to modal logic, London, UK: Routledge.

MCCARTHY (1990). Elephant 2000: A programming language based on speech acts, Unpublished manuscript.

SEARLE, J.R. (1969). Speech Acts, Cambridge, UK: Cambridge University Press.

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, pp. 344–369. Reprinted in J. R. SEARLE (1979). Expression and meaning, Cambridge, UK: Cambridge University Press.

SEARLE, J.R. AND VANDERVEKEN D. (1985). Foundations of illocutionary logic, Cambridge, UK: Cambridge University Press.

SHOHAM, Y. (1993). Agent-oriented programming. *Artificial Intelligence* 60: 51–92.

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

WOOLDRIDGE, M. (2000). Reasoning about rational agents, Cambridge, MA: MIT Press.