Sarcasm, deception, and stating the obvious: planning dialogue without speech acts

Debora Field*
(d.field@umist.ac.uk)
Dept. of Language and Linguistics, UMIST, PO Box 88, Manchester M60 1QD, UK

Allan Ramsay†
(allan@co.umist.ac.uk)
Dept. of Computation, UMIST, PO Box 88, Manchester M60 1QD, UK

Abstract. This paper presents an alternative to the 'speech acts with STRIPS' approach to implementing dialogue: a fully implemented AI planner which generates and analyses the semantics of utterances using a **single linguistic act** for all contexts. Using this act, the planner can model problematic conversational situations, including felicitous and infelicitous instances of bluffing, lying, sarcasm, and stating the obvious. The act has negligible effects, and its precondition can always be proved. 'Speaker maxims' enable the speaker to plan to deceive, as well as to generate implicatures, while 'hearer maxims' enable the hearer to recognise deceptions, and interpret implicatures. The planner proceeds by achieving parts of the constructive proof of a goal. It incorporates an epistemic theorem prover, which embodies a deduction model of belief, and a constructive logic.

Keywords: AI planning, conversational record, deception, implicature, speech acts

1. Introduction

Example (1)

Initial state	John has been bird-watching with Sally for hours, and so far, they have only seen pigeons. John thinks Sally is feeling bored and fed up. John has some chocolate in his bag.
Goal condition	John wants to cheer Sally up.
Solutions	John is just thinking about getting out some chocolate to give her, when yet another pigeon lands in a nearby tree. John sees an opportunity to make Sally laugh by means of a bit of sarcasm, and so plans to say to her, **"There's an albatross!"**

[*] Tel: + 44 161 200 3120; Fax: + 44 161 200 3091

[†] Tel: + 44 161 200 3108; Fax: + 44 161 200 3324

This paper presents a fully implemented model of dialogue in which the semantics of utterances are generated using a single linguistic act for all contexts.[1] The situation described by Example (1) is typical of the planner's tasks.[2] The tasks depict agent John in social situations with one other named agent, and having an intention to affect the other agent's belief state in some particular way. When a task is input, the planner plans linguistic and non-linguistic actions to achieve John's intentions. In Example (1), John's intention is to cheer Sally up. The first plan he makes is to give her some of the chocolate he has brought along:

```
P = [(mech,{give,john,to,sally,chocolate_bar})] ?
```

. . . and the second plan he makes is to tell her that there is an albatross:[3]

```
P = [(ling,{john,to,[sally],species(bird_1,albatross)})] ?
```

John uses conversational maxims to plan utterances, which enable him to generate implicatures, and plan deceptions. The reason John decides to tell Sally that there is an albatross is because he thinks telling her this constitutes a joke—doing this flouts (blatantly contravenes) a particular maxim, and floutings of this maxim indicate, among other things, humour. John plans one or two utterances without deviating from Grice's Cooperative Principle, but for most of the tasks, he contravenes the CP by flouting or violating a Gricean maxim.

In order to model felicitous and infelicitous hearer responses, the belief state of John's hearer is included in the formal task definitions. (Although Example (1) does not mention it, each code version of Example (1) contains a belief state for Sally as well as John.) The hearer uses conversational maxims to interpret the speaker's utterances. These are similar to the speaker's maxims, but are designed for the purposes of **recovering** implicatures, and **recognising** attempts to deceive.

Once John has planned a linguistic action, the action is applied to John's initial belief state and to the hearer's initial belief state, in order to derive new post-utterance belief states for each of them, the states that would be attained if John's planned meaning were given expression in that particular situation. From these states, **we** can see

[1] The term 'linguistic act' in this paper denotes a formal STRIPS-style (Fikes and Nilsson, 1971) action operator.

[2] Example (1) is an English paraphrase of a task, written in the model in Prolog code. Section 7.2 contains examples of code from a task

[3] To assist understanding of the examples, some zoological clarification will be included in the footnotes. An albatross (*Diomedea exulans*) is an enormous sea-faring bird, rarely seen from the land.

whether John's utterance would have been felicitous or not, however, John cannot, since John does not know and cannot observe his hearer's actual belief state.

Whether John's intentions would be achieved by his utterances depends on *H* having the 'right' set of beliefs (the ones John thinks she has) and making the 'right' inferences (the ones John expects her to make). John plans (the semantics of) an utterance, expecting that the utterance would have a particular effect on *H*'s belief state; if John were to perform the utterance, he would not be certain that it had achieved his intention, but he would expect that it probably had.

For example, if John's utterance "There's an albatross!" in Example (1) is to be felicitous, the following must happen. Sally must first believe that John has said something that Sally thinks John and Sally mutually believe is false. From this, she must infer that John has flouted a conversational maxim, and consequently that John has attempted to implicate a meaning which is not expressed by the semantics of "There's an albatross!". Sally must then infer from this flouting that John has said "There's an albatross!" in order to try and cheer her up by making a joke. Whether or not any of this happens depends on Sally's belief state, (which John cannot observe, but we can).

2. Motivation: Problems with speech acts

The speech acts approach to natural language (NL) processing envisages a selection of different acts from which the speaker chooses the act that suits his particular beliefs and intentions in the given situation. The approach is inspired by Searle's (1965, pp. 46-47) "illocutionary act of promising", which has nine conditions "necessary and sufficient for the act of promising to have been performed in the utterance of a given sentence". These nine conditions, seven of which do not concern surface linguistic form, legislate for the intentions of the speaker, and for the beliefs of both the speaker and the hearer.

In the early days of AI planning, some researchers saw that marrying speech act theory and AI planning together could have great potential for NL generation, and they began to represent speech acts using the STRIPS (Fikes and Nilsson, 1971) operator (Bruce, 1975; Cohen and Perrault, 1979; Allen and Perrault, 1980; Appelt, 1985). Different angles on the 'speech acts with STRIPS' approach have been tried, but the approach is fraught with well-documented difficulties (Cohen and Levesque, 1990; Grosz and Sidner, 1990; Pollack, 1990; Bunt, 2000; Ramsay, 2000), all of which relate to the inability of participants in a conversation to observe each other's beliefs and intentions. The research leading to this paper was carried out in an attempt to implement an alternative to the speech acts approach to NL generation.

3. Aim: To plan meanings

The planner presented here models the communication of meanings between agents situated in contexts. In the real world, when considering how to express his intended meaning, a person not only has words (and other linguistic signs) at his disposal, but also means of a non-linguistic kind. For example, he may pull an unhappy face to communicate dissatisfaction, or eat meat to communicate that he is not a vegetarian, or say to someone, “My name is John”, to inform her that his name is John. As there are very many physical actions (and combinations thereof) that can be used by a person to communicate information, there is, then, an unspecifiable number of what we might call ‘communicative actions’. The planner presented here plans the **semantics** of such actions, and it refers to them not as ‘communicative actions’, but as ‘linguistic actions’.

This planner does not plan the surface realisations of linguistic actions. By ‘surface realisations’ is meant both the symbols that are often used in physical expressions of meaning, such as phonemes, or hand shapes, and the actions of physical expression themselves, such as changing one’s facial muscles, or writing. In the model, in order for a planned linguistic action to have the speaker’s intended effect on a hearer’s belief state, some means of physically expressing the planned meaning by some symbolic means would also have to be planned, and performed. The model’s planned linguistic actions are, however, often discussed **as if** they had been given a form of physical expression, in which cases, successful, unimpeded production (by S) and perception (by H) of suitable symbols is assumed. No assumption or claim is being made that planning the surface realisation of an intended meaning, and planning the intended meaning itself, are separate processes, nor is the opposite being claimed.[4]

4. Development process

The original aim of this research was to develop a small set of linguistic acts that were unambiguously identifiable purely by surface linguistic form (after Bunt 2000), including ‘declare’, ‘request’, and perhaps

[4] Investigation of the relationship between a planned meaning and its surface realisation is an obvious direction for future research.

others—a set of acts with negligible effects (after Ramsay 2000), and minimal preconditions. The experiment began by designing a STRIPS operator to capture the notion of informing. 'To inform' was simplistically interpreted as 'to make know', the idea being that the effect of informing some agent A of a proposition P is *knows(A,P)*. The operator's design was then challenged with some problematic conversational scenarios, of which Example (1) is one, and which focused on instances of sarcasm, stating the obvious, bluffing, and lying. Questions which arose during this process, and which shaped the emerging form of what was to become the linguistic act, were:

Preconditions

Given that people have the physical and mental capabilities to say whatever they like at any moment they choose, what inhibits them from doing so, and is it anything to do with the preconditions of linguistic actions?

Effects

Given that by performing an utterance, the speaker (S) does not directly influence H's belief state, but has to rely on suitable hearer response for speaker meaning to be understood, what is the relevance of this to the effects of formal linguistic act operators? **Can** linguistic actions be represented by STRIPS-like operators, and if yes, what are their effects lists like? How does the application of a speaker's linguistic action affect his own beliefs about H?

Deception

Given that speakers often violate maxims, and assuming that it is in the interests of H to detect maxim violations,[5] what strategies do hearers use to discover whether or not they have been deceived? How can a speaker plan to deceive, in other words, how can a speaker intend that H not detect that he is attempting to deceive her?

Implicature

Given that the surface form of a speaker's utterance often does not encode S's meaning, how does the semantics of what he says relate to his intended meaning, and how can H decipher S's intended message?

[5] That it is in the interests of H to detect deceptions is not always the case in the real world, as S may be lying to protect to H in some way. This model, however, assumes that that, even though he knows this is a possibility, a hearer nevertheless does not want to be deceived.

Collaboration

Given that S can lie to H, and given that H can pretend to believe S while actually disbelieving him, how can such complex conversations continue without descending into confusion after the first utterance?

After much experimentation, a model was developed which relies on only one linguistic act, rather than on the small set of acts which had been anticipated. The single linguistic act carries no illocutionary force or perlocutionary intention, and in this respect is after Austin's locutionary act (Austin, 1962, p. 109). Vital additional mechanisms work alongside the linguistic act, inspiration for which was taken from previous work concerning Lewis's 'accommodation' (Lewis, 1979), the 'conversational record' (Stalnaker, 1972; Thomason, 1990), and Grice's CP and conversational maxims (Grice, 1975).

5. Planner design

5.1. Planning approach

The ultimate application of the planner, to plan linguistic actions, strongly influenced decisions that were made early on in the research concerning planning approach. The approach of many of today's domain-independent planners is to use heuristics either to constrain the generation of a search space, or to guide the search through the state-space for a solution, or both (Blum and Furst, 1995; Bonet and Geffner, 1998; Hoffman and Nebel, 2001; Nguyen and Kambhampati, 2001). Although it was anticipated that some kind of search-constraining heuristics would be needed for this planner, it was clear that a model of human dialogue planning would have to be able to do more than this. The planning agent would need to do reasoning about its current situation, and its goals, in order to decide how to act. He would also have to be able to analyse actions in their wider setting, to understand that his linguistic actions could have many different effects, depending on the context, and to be able to plan utterances in spite of this.

There was also the question of search direction. Some of the most successful planners today search forward from the initial state. In the early stages of the research, it was not known what the preconditions of linguistic action operators would be like, however, it was nevertheless considered that a forward planning approach would be highly unsuitable. People generally have the physical and mental capabilities to say whatever they want at any moment. This meant the answer to the question 'What can I say in the current state?' was anticipated to be something like 'Anything, I just have to decide what I want to say'. It was therefore considered that a backward search, which concentrates on an agent's desires, would be far more suitable than a forward search, which concentrates on an agent's capabilities.

A planner was designed that reasons with its beliefs about how to act in order to achieve its goals, and that can achieve a goal that does not match any action effect, but that is entailed by the final state. It does this by achieving parts of the constructive proof of that goal via an interweaving of reasoning, backwards state-space search via STRIPS operators, and some specially designed techniques and heuristics. The final model was developed by first implementing (in Prolog) a STRIPS-style planner based on foundational work in classical planning (Newell and Shaw and Simon, 1957; Newell and Simon, 1963; Green, 1969; McCarthy and Hayes, 1969; Fikes and Nilsson, 1971). A theorem prover for first-order logic was then implemented, developed into an epistemic theorem prover, and then incorporated into the planner.

5.2. Epistemic theorem prover

The epistemic theorem prover has three distinctive features: it embodies a constructive/intuitionist logic (developed by Brouwer (1881-1966) (Heyting, 1956; Heyting, 1975)), it proves theorems by natural deduction (Gentzen, 1935) and it embodies a deduction theory of belief (Konolige, 1986).

5.2.1. *Constructive logic and natural deduction*

Using constructive logic, and doing natural deduction, provides a more accurate representation of human reasoning than using classical logic. When reasoning about anything, including about how to act, humans are interested in the content of an argument; they see content relationships between the conclusions that are drawn, and the premises from which they are drawn. **How** a proposition P is concluded, from which premises, and using which rules, is the focus of attention. In contrast, the emphasis in classical logic is on the question, '**Can** we prove P?'. Classical logic will tell you whether P is true or false. Constructive logic will construct an argument which proves P, or it will fail.

Understanding of the meaning of a proof is, then, of little interest in classical logic. This is reflected in the meaning of the implication operator (here written as $\Rightarrow$), which is fully expressed by the truth table for $\Rightarrow$. In constructive logic, however, the way in which the operator $\Rightarrow$ is used hints at a relationship between antecedent and consequent—there is something about P being true that is in some way connected to Q being true. This interpretation of $\Rightarrow$ is much closer to the human understanding and use of terms like 'implies', or 'if…then' than the classical interpretation.

5.2.2. *Deduction model of belief*

Konolige's deduction model of belief was chosen because it was considered far more representative of human reasoning than the 'possible worlds' model (Hintikka, 1962; Kripke, 1963), a popular choice in AI since Moore's implementation of it (Moore, 1985). In a possible worlds model, an agent must believe absolutely every proposition that follows from his belief set; he is a supremely powerful and ideal reasoner. People, in contrast, are not ideal reasoners; their belief states are incomplete, inaccurate, and inconsistent, and their inference strategies are imperfect. The agents in Konolige's model are imperfect reasoners: an agent does not have a complete set of inference rules, and is therefore not forced to infer every proposition that follows from his belief set.

Additionally, an agent who is logically omniscient uses the implication operator in a different way to humans: knowing Q automatically follows from knowing $P \wedge (P \Rightarrow Q)$; there is no progressive 'drawing' of the conclusion Q, no awareness by the agent that the reason why he knows Q is because he knows $P \wedge (P \Rightarrow Q)$. Konolige's deduction model avoids the problem of logical omniscience, enabling the **process** of inference to be modelled (**how** P is concluded, from which premises, and using which rules). It enables representation of what an agent **actually** believes, rather than of all the possible beliefs that an agent could, hypothetically, derive from his belief state.

6. Design of the linguistic act

6.1. The effects of the linguistic act

When S plans to utter a proposition to H, he cannot know what H will believe as a consequence of his utterance. Even after the utterance has been executed, its effect on H is unobservable by S; and even if H does infer S's intended meaning, S cannot know whether his intended meaning is *believed* by H. S is totally reliant on the intention of H to infer the meaning that she chooses to infer from S's utterance—there is no (cognitive) effect on H which does not originate in H's intention. Consider this example:

Example (2)

Initial state	John has gone bird-watching with Sally. John likes Sally a lot, and is enjoying telling her about birds. He thinks Sally is new to bird-watching. He also thinks that she looks cold.
Goal condition	John wants Sally to be impressed by him.
Solutions	John is just thinking of offering Sally his coat to wear, when a huge bird lands in a nearby tree. John isn't quite sure what species the bird is, nevertheless, he decides to try and impress Sally with his bird expertise, and plans to say, **"There's a dodo!"**

If John were to express this utterance to Sally, whether or not John would be successful in concealing his bluff, and therefore in impressing Sally with his identification of a dodo,[6] would depend on whether his beliefs about Sally were correct. For example, whereas John may think that Sally is less knowledgable about birds than he is, he may be mistaken. In one of the model's tasks, Sally sees through John's bluff, because she is quite sure that what they are looking at is a buzzard.[7] However, John cannot predict whether Sally will believe his utterance, and infer from it that John is an impressive chap, or whether she will see through his attempted bluff, and consequently be very unimpressed.

Since linguistic actions have neither predictable nor observable effects, the authors consider it a waste of effort categorising linguistic actions according to effect by having a library of different acts with different effects—a speaker will never know whether the act he thinks he has performed has been performed, so why should he bother worrying about it? What **is** important to S is having an expectation **that *H* will believe something particular** after hearing S's utterance, not an expectation **that the effects of a particular linguistic act will be achieved** by the utterance.

A model which is not compromised by S's inability to know or observe the effects of his utterances is one in which all linguistic actions **have the same effect**. Such a model is presented here, one in which

[6] A dodo is a large flightless bird that is famously extinct.

[7] A buzzard (*Buteo buteo*) is a large bird of prey with a powerful hooked bill and sharp talons.

every linguistic action has the effect of **recording a new entry in the minutes**. When a plan to utter P is applied to a current state, a minute of the form *minute(S,H,P)* is added to the current state. From the new state with added minute, S and H make inferences which transform their pre-utterance belief states into new belief states.

6.1.1. *The conversational 'minutes'*

The minutes are a variation on the idea of the conversational record (Stalnaker, 1972; Lewis, 1979; Thomason, 1990). To S and H, the minutes represent the semantics of an 'ideal' conversation, the conversation that would have emerged under 'ideal' circumstances. These circumstances include, for example: H has not misheard S; S and H each believe each minute is true; the things that S assumed that H already believed, H did in fact already believe; H associates the referring expressions in the minute with the same referents that S does; and H resolves any unintended ambiguities in the 'right' way. From his view of the minutes, and his (private) belief state, each conversant makes inferences that transform his pre-utterance belief state into a new belief state.

6.1.2. *The speaker's goal*

By proposing that the effect of all linguistic actions is the recording of a proposition in the minutes, it is not being suggested that all linguistic actions have the same meaning, or the same goal, or that linguistic actions do not have to be planned. Achieving the recording of a new minute clearly does not fully embody S's intention, since the reason he planned an utterance at all was because he wanted to change H's belief state in a particular way. However, although S cannot **know** what H will believe as a consequence of S's new entry in the minutes, he generally **expects** that his desired change to H's beliefs will occur, otherwise he would not bother to speak.

6.2. The preconditions of the linguistic act

In AI planning theory, the preconditions of all actions, including linguistic ones, are generally considered to be those conditions which must be true in the current state for an action to be possible in the current state. In order to make a plan that would result in the existence of a fresh cup of hot coffee, one has to reason about practical matters, such as whether or not one has a kettle, a cup, water, and so on; these practical details can be thought of as preconditions of making a coffee.

In order to make a plan that would result in the existence of a meaning in one's own mind, one does not need to reason about practical matters at all, such as whether or not one has a kettle, or anything else. So do linguistic actions have preconditions at all?

People often say things they think their hearers do not know, in order to warn them they are in danger, perhaps, or to let them know they will not be home for dinner. People also say things that they think are already mutually believed by themselves and their hearers, in order to convey surprise, perhaps, or boredom. People even say things they think are mutually believed to be **false** by themselves and their hearers, in order to make a joke, perhaps, or to convey disapproval. What is more, people also say things they themselves believe are false, or believe are true, or are not sure about. An agent who is planning the meaning of an utterance can clearly plan whatever meaning he chooses to, regardless of his beliefs.

So, people have the mental capabilities to plan whichever meanings they want to at any and every moment, and, assuming normal physical capabilities, to communicate (via speech, gesture, hand signs,...) whatever they like at any and every moment—**but they do not do this**. There are many pragmatic reasons why this may be the case (turn-taking, politenesss, consideration,...), however, we argue that underlying them all is a fundamental limitation: people are only able to plan and express meanings **of which they are aware in the current moment**. Otherwise, people **cannot** plan or express meanings.

People can in certain circumstances utter propositions which do not mean anything to them. However, if a person utters a proposition, but is unaware of the meaning of that proposition, he must either have planned a **physical** action, (for example, he may have repeated a phrase in a language he does not understand), or he must be somehow disconnected from his normal rational thought processes, (perhaps due to mental illness, for example).

These ideas are embodied by the single disjunctive precondition of the linguistic act:

```
believes(S, P)
or believes(S, not(P))
or believes(S, (P or not(P)))
```

The expression *believes(S,P)* is best understood as a **current active believing of *P*** by *S*, rather than the holding of a belief by *S* of which he is not currently aware. Before *P* is brought to *S*'s attention, *S* is not aware of his beliefs about *P*, and so *not(believes(S, P))* holds. *S* becomes aware of his view of P by being prompted, either by someone else's utterance, or an appetite, or a desire, and so on; S's absence of a view of P is transformed into an actual view of P in that moment, making him momentarily aware of something new.

7. Modelling deception

John and his hearers conduct conversation according to Grice's Cooperative Principle, which Grice breaks down into specific maxims (Grice, 1975). Here are two of them (*ibid* p. 308):

> "[**Quantity**]
>
> 1. Make your contribution as informative as is required (for the current purposes of the exchange).
> 2. Do not make your contribution more informative than is required...
>
> [**Quality**]
>
> 1. Do not say what you believe to be false.
> 2. Do not say that for which you lack adequate evidence."

The Quantity maxims are concerned with what S thinks H needs/wants to know, whereas the Quality maxims are concerned with what S himself believes. Grice's maxims prescribe a standard for speaker behaviour which S can then blatantly contravene ('flout'), thus signalling to H that there is an implicature to be recovered.

Grice's maxims would be adequate for modelling H's understanding of an utterance in a world in which no-one ever tried to deceive anyone else. However, as we know, people **violate** maxims—they contravene maxims without wanting their hearers to know. This means that when a hearer is interpreting the meaning of a speaker's utterance, she must take into account the possibility that S is trying to deceive her. This in turn means that to model the planning of deceptive utterances, H needs two kinds of maxims: (i) maxims which embody Grice's CP, *i.e.*, the standard which S is **supposed** to adhere to; and (ii) some more practical maxims which take account of the fact that speakers do not necessarily always adhere to the CP.

7.1. Hearer maxims

Grice's maxims of Quantity and Quality are adopted in this model as the first kind of maxims H needs, the ideal standard for conversation, and these are called here the 'general maxims'. When H assesses a speaker's utterance, she considers that it is a possibility that S has not adhered to the general maxims, and on these grounds, she employs a number of more practical maxims, which will be called 'hearer maxims' (or 'H_maxims').

Given that H admits the possibility that S might be trying to deceive her with regard to his utterance, we consider that there are two strong predictors of how H will respond to a speaker's utterance, in addition to her pre-utterance view of the proposition P that has been uttered. H's response, then, will depend on her precise answers to the following three questions:

i **What is *H*'s view of the proposition *P*?**
ii **What is *H*'s view concerning the goodwill of *S*?**
iii **What is *H*'s view of the reliability of *S*'s testimony?**

For example, in an infelicitous bluff task (based on Example (2)), Sally's answers to these questions are as follows. Before John performed his utterance:

i Sally believed that the proposition P ("There's a dodo!") was false (because she knew the bird was a buzzard).
She did not believe that John thought that they mutually believed P was false.
ii She believed that John was well-disposed towards her.
iii She didn't know whether John was a reliable source of information or not.

After John has said "There's a dodo!", Sally derives the following new set of beliefs from the above set:

i′ Sally still believes that the proposition P ("There's a dodo!") is false.
She **now** believes that John thinks that they mutually believe P is true.
ii′ She still believes that John is well-disposed towards her.
iii′ She **now** believes John is an unreliable source of information.

This pattern of belief transformation is determined by a hearer maxim which prescribes that the hearer should respond in accordance with the second belief template above, if presented with a situation in which the first belief template applies; we might call this maxim the 'infelicitous bluff' H_maxim.

The model currently has twelve different H_maxims. Eight of the H_maxims define particular changes to H's view, concerning the proposition uttered, the reliability of the speaker, and the goodwill of the speaker, under particular conditions. In addition to the 'infelicitous bluff' maxim illustrated, there is an 'infelicitous deceive' H_maxim (which prescribes that *H*'s view of *S*'s goodwill is brought into question), as well as some 'infelicitous inform' maxims, a 'felicitous inform' maxim, and others. Two of the H_maxims define how *H* recognises the generation of an implicature; she infers an implicature if the proposition *S* has uttered is one she thinks they already mutually believe or disbelieve, regardless of her view of *S*'s goodwill and reliability.

7.2. Speaker maxims

In a conversation in which deception is believed to be a likelihood, *S* and *H* can be pictured as opponents in a battle of wits. Each is trying to thwart the efforts of the other: *H* is trying to avoid being deceived by *S*, whereas *S* is trying to avoid *H* detecting his intention to deceive her. If *H* is going to avoid being deceived, she will have to consider not only what *S* says, but also what she thinks about *S*. Therefore, if *S* is going to succeed in his deception, he will have to take into account how *H* is going to try and detect his deception. To represent this in the model, *S* has his own 'speaker maxims', which concern the same issues as the H_maxims, but from the other side of the table, as it were. What *S* plans to say will depend on which answer he selects from each of these four categories:

i **What is *S*'s view of *H*'s view of various different propositions?**

ii **What is *S*'s own view of the same propositions?**

iii **What is *S*'s view of *H*'s view of the goodwill of *S*?**

iv **What is *S*'s view of *H*'s view of the reliability of *S* as a source?**

The model currently has five S_maxims which prescribe what *S* can expect *H* will believe, if *S* performs utterances under particular conditions. Here are two examples of the S_maxims (paraphrase first):

S_maxim 1: Informing and lying

If I put Q into the minutes under these conditions, (even if I don't believe Q), then I believe H will believe I am being Grice-cooperative[8] with respect to Q:

```
minute([S], [H], Q)
    and believes(S, believes(H, reliable(S)))
    and believes(S, believes(H, well_disposed_towards(S, [H])))
    and believes(S, believes(H, Q or not(Q)))
==> believes(S, believes(H, gricecoop(S, [H], Q)))
```

S_maxim 2: Stating the obvious

If I put Q into the minutes under these conditions, then I believe H will believe I am being Grice-uncooperative with respect to Q:

```
minute([S], [H], Q)
    and believes(S, believes(H, mutuallybelieve(([H, S]), Q)))
==> believes(S, believes(H, griceuncoop(S, [H], Q)))
```

The consequent of S_maxim 1 is that S believes H believes S is being **Grice-cooperative** about Q. To plan a deception, S reasons that if he plans a linguistic action according to the specified conditions, then H will believe S is being Grice-cooperative, and H will believe that what S is saying is true (even though S does not believe it himself). If S uses this S_maxim to plan a deception, he will violate a **general** maxim (Quality 1), which asserts that people should not say what they believe to be false.

The consequent of S_maxim 2 is that S believes H believes S is being **Grice-uncooperative**, in that he is flouting a maxim. (S is not planning that H infer S is trying to deceive her (although a maxim could easily be added to represent this complicated kind of intention)).

7.2.1. *Which meanings can be implicated?*

The S_maxims are not enough on their own to enable John to plan utterances which will achieve his intentions. John also has to have views about what H will infer from his use of the S_maxims. Some of the opinions John has about what H will infer are as follows. Each inference rule is preceded by an inaccurate but informative enough

[8] The term 'Grice-cooperative' here means 'abiding by the general maxims' as outlined in 7.

English summary:

Grice-coop implies H will believe what I say

```
believes(S,believes(H,
   gricecoop(S, [H], Q) ==> Q))
```

Grice-uncoop implies H will infer I'm bored

```
% John believes that Sally believes that if P2 is being
% Grice-uncooperative about some Q which P2 and P1 have a
% mutual belief about then maybe P2 is bored by Q.
believes(john,believes(sally,
   (griceuncoop(PERSON2, PERSON1, Q)
        and mutuallybelieve(([PERSON2, PERSON1]), Q)  )
        ==> bored(PERSON2, re(Q))  ))
```

7.2.2. *Speaker and hearer maxims: A detailed example*

Here is a detailed example of how the S_maxims are used to plan an utterance, and how the H_maxims are used to interpret it.

Example (3)

Initial state	John's friend Andy has joined John and Sally in the forest to do some bird-watching. John likes Sally a lot and doesn't want Andy competing with him for Sally's attentions. John knows that Sally is single.
Goal condition	John wants to discourage Andy from pursuing Sally.
Solutions	John therefore decides to say, **"Sally is married."**

Here is the goal condition as expressed in the model (in Prolog code):

```
believes(john,believes(andy,
    discouraged(andy, re(sally))))
```

. . . , and here is the code for John's initial (pre-utterance) belief state:

```
believes(john, not(married(sally))),

believes(john,believes(andy,
    (married(sally) or not(married(sally))))),
believes(john,believes(andy,
    gricecoop(_PERSON1, _PERSON2, married(sally))
        ==> married(sally)   )),

believes(john,believes(andy,
    married(sally) ==> discouraged(andy, re(sally)) )),

believes(john,believes(andy,
    reliable(john))),

believes(john,believes(andy,
    well_disposed_towards(john,[andy]))),

believes(john, not(well_disposed_towards(john,[andy]))),
```

In order to achieve his goal, John plans to put "Sally is married" into the minutes using the 'informing and lying' S_maxim (see 7.2). In using this S_maxim, John lies, and thus violates a general maxim (Quality maxim 1), because he knows that Sally is single. John thinks his deception will succeed, because he thinks that the conditions of the S_maxim hold. John thinks that: (i) Andy thinks that he (John) is a reliable source of information; (ii) Andy thinks that he (John) is well-disposed towards Andy (even though, in fact, John is not); (iii) Andy doesn't know whether Sally is married or not. John is right about the first two conditions, but he is wrong about the third.

Whether or not John's goal would be achieved by his utterance depends on Andy's pre-utterance beliefs about Sally and John, and on Andy's moral stance on extra-marital affairs. This is Andy's pre-utterance belief state:

```
believes(andy, reliable(john)),

believes(andy, well_disposed_towards(john, [andy])),

believes(andy, not(married(sally))),

believes(andy, married(X)) ==> discouraged(andy, re(X)),
```

Andy does believe that John is a reliable source of information, and that John is well-disposed towards him. However, Andy knows more about Sally than John is aware of, and already knows that she is not married. This means that, whereas John's intention is that Andy should infer John is being Grice-cooperative, Andy does not infer this: John has said a proposition P that Andy knows is false, but P is not something that Andy thinks they mutually believe is false, and so Andy infers that John has violated a general maxim (Quality 1, the one that John has in fact violated). This means that Andy does not add P to his own private beliefs, however, he does add to his beliefs that **John** believes that Andy has added P to his (Andy's) beliefs, and Andy also changes his opinion about the goodwill of John towards him.

During application of John's planned linguistic action to the initial state (as if it had been given physical expression), one change is made to John's pre-utterance belief state, apart from the addition of the minute *minute(john,andy,married(sally))*. John's belief that Andy doesn't know whether Sally is married is changed to the following belief:

```
believes(john, believes(andy,
    mutuallybelieve(([andy,john]),married(sally))))
```

As John expects that his lie has not been detected, he expects that Andy now believes that he (Andy) and John mutually believe Sally is married—however, John himself still believes that Sally is not married.

The changes that Andy makes to his pre-utterance belief state as a consequence of John's utterance (apart from the addition of the minute) include the following additions:

```
believes(andy, (griceuncoop(john, andy, married(sally)))),

believes(andy, not(well_disposed_towards(john,[andy]))),

believes(andy, believes(john,
    mutuallybelieve(([andy,john]), married(sally)))),
```

Part of Andy's inferring that John has lied is inferring that John now believes that Andy and John believe that they mutually believe Sally is married—however, Andy himself still believes that Sally is not married.

8. Yes/No questions

Some experiments were carried out on modelling yes/no questions by designing some 'verification tasks' which employ an interrogative version of S_maxim 1 ('informing and lying', see 7.2). In one verification task John finds himself having to deal with a live bomb. John knows that he can disarm the bomb by cutting one of the wires, but he does not know which wire to cut to avoid setting off an explosion. John, however, thinks that Dave will know, so John plans to contact Dave, and ask him. The goal for this task is *believes(john,safe(bomb_1))*, and the solution returned for the task is:

```
P = [(ling,{john,to,[dave],
       (earthed(wire_1) or not(earthed(wire_1)))}),
     (ling,{dave,to,[john],
        earthed(wire_1)}),
     (mech,{[john],disarm,bomb_1})] ?
```

A bomb disposal situation was chosen for the verification tasks to emphasise that if John were to cut the wire having made this plan, but not having given it some form of expression, **and** having assessed the response, then he would be in grave danger. In order for John to know whether the wire is in fact earthed, he will have to give physical expression to his plan; Dave's answer, that the wire is earthed, is not being predicted by John, John is reasoning that if Dave told him this, he would believe Dave, and then it would be safe to disarm the bomb. Further development of the work on yes/no questions is a clear avenue for future research, as is the implementation of other classes of utterance.

9. Summary

The model presented by this paper is one in which: all linguistic actions have the same effect (the recording of a proposition in the conversational minutes); all linguistic actions have the same disjunctive precondition, one which, when verified by S, has the significant side-effect of making S aware of something new; and in which conversants use Gricean maxims as a standard they expect others to abide by, as well as more practical maxims which enable them to generate and recognise violations of Grice's CP. Using a single linguistic act for all contexts, the fully implemented planner can plan the semantics of utterances for a range of difficult conversational situations, including situations in which an agent will make a plan to communicate:

things that he believes H does not know, and either needs or would want to know (for example, a warning of danger, or an alert to new information);

things he is quite sure other participants in the conversation already believe (for example, to express surprise or boredom);

things that he thinks are blatantly false, for the sake of sarcasm (for example, for the purposes of humour, or criticism);

things that he believes are false, but that, for personal reasons, he wants to be dishonest about (for example, for laudable reasons, such as protecting the innocence of a child, and for less laudable ones, such as protecting his own romantic interests);

things that he is not sure about, but that he wants to pretend he knows (for example, to impress H);

things that he is not sure about, and wants to be sure about, and that he thinks H will know (for example, a question to elicit a response from H).

The model also has tasks in which it can be seen whether S's (John's) intention would be achieved by his plan, if it were to be expressed to H. These include situations in which H:

correctly infers that John is intending to deceive her;

fails to infer that John is intending to deceive her;

doesn't know whether John is a reliable source of information or not, but changes her view to believing that John is an unreliable source, after inferring either that he is being Grice-cooperative, but ignorant, or that he is trying to deceive her, by informing her of something she already knows is incorrect;

believes that John well-disposed towards her, and either believes John is not a reliable source of information, or doesn't know whether he is or not, but changes her view to believe that he is a reliable source, after observing him try to inform her of something she already knows is correct;

believes John is a reliable source of information, and is well-disposed towards her, and so infers that John is being Grice-cooperative in trying to inform her of something, and gains a view of John's expressed proposition which is the same as John's expressed view, because she has no reason to do otherwise;

believes John is not a reliable source of information, but that he is well-disposed towards her, and so infers that John is being Grice-cooperative in trying to inform her of something, but gains a view of John's expressed proposition which is not the same as John's expressed view, because she has reasons for not adopting his view;

correctly recognises John's signal that an implicature has been generated, and decodes it according to John's intention;

correctly recognises John's signal that an implicature has been generated, but does not decode it according to John's intention;

fails to recognise John's signal that an implicature has been generated, and so does not infer John's intended meaning.

Additional work has been carried out on modelling yes/no questions, and some tentative experiments have also been carried out concerning planning the physical symbolic expression of meaning.

10. Discussion

After a person has made an utterance, there is normally little doubt among conversants that some particular sounds have been uttered, or visual signs made, and that **something** has been added to the minutes, however, there is no guarantee that S and H concur on **what** has been added. Consider a situation in which a speaker utters "There's a kite!" to a hearer, having seen a bird on a fence. In so doing, the speaker has put the proposition *species(object_1,kite)* into the minutes, referring to the bird on the fence, and using the term 'kite' to mean a particular species of bird. The meaning of "There's a kite!" to the hearer, however, may differ. The hearer may not know that there is a species of bird called 'kite', and may think the speaker is talking about a toy. Alternatively, the hearer may know that there is a bird called a 'kite', but may think that the speaker is referring to a different bird, not the bird on the fence.

It is therefore considered that in an ideal model, S and H would have their own private beliefs concerning the contents of the minutes. Currently, this implementation does not represent this. As a consequence, misunderstandings over referents cannot be modelled by the current implementation, nor can any other forms of misunderstanding which arise out of ambiguities—an assumption is made that S and H concur on all these things. Implementing private views of the minutes would be a straightforward and highly desirable improvement to the model.

The current modelling of the reliability of an agent is inadequate, and requires a finer-grained representation. People do not normally consider other people to be unreliable sources of information on all subjects, they usually consider that other people can provide reliable information on **some** subjects, but not on others.

Other directions for future work include:

- development of planning processes to decide on maxims and implicatures;
- experimentation with making the decision to flout a maxim a default decision;
- development of more speaker (and hearer) maxims,*e.g.* a 'being economical with the truth' maxim and a 'deliberate confusion by too much information' maxim;
- implementation of tone of voice as an aid to implicature recovery;
- implementation of degrees of belief.

Acknowledgements

This research was funded by an EPSRC studentship.

References

Allen, J. F. and C. R. Perrault. 'Analyzing intention in utterances'. *Artificial Intelligence* 15:143–78, 1980.

Appelt, D. E. 'Planning English referring expressions'. *Artificial Intelligence* 26:1–33, 1985.

Austin, J. L. *How to do things with words*, 2nd Edition. Oxford: OUP, 1962.

Blum, A. L. and M. L. Furst 'Fast planning through planning graph analysis'. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1636–1642, 1995. Also in *Artificial Intelligence* 90: 281–300, 1997.

Bonet, B. and H. Geffner 'HSP: Heuristic Search Planner'. *AI Magazine* 21(2), 2000.

Bruce, B. C. 'Generation as a social action'. In B. L. Nash-Webber and R. C. Schank (eds.), *Theoretical issues in natural language processing*, pp. 74–7, Cambridge, MA: Ass. for Computational Linguistics, 1975.

H. Bunt 'Dialogue and context specification'. In H. Bunt and W. Black (eds.), *Abduction, belief and context in dialogue: Studies in computational pragmatics*, pp. 81–150, Philadelphia, PA: John Benjamins, 2000.

Cohen, P. R. and H. J. Levesque 'Rational interaction as the basis for communication'. In Cohen et al., pp. 221–55, 1990.

Cohen, P. R., J. Morgan and M. E. Pollack (eds.) *Intentions in communication*. Cambridge, MA: MIT, 1990.

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

Fikes, R. E. and N. J. Nilsson 'STRIPS: A new approach to the application of theorem proving to problem solving'. *Artificial Intelligence* 2:189–208, 1971.

Gentzen, G. Investigation into logical deduction. In M. E. Szabo (ed.), *The collected papers of Gerhard Gentzen*, pp. 68–131. North-Holland, 1969. Originally published as 'Untersuchungen über das logische Schliessen', 1935, in *Mathematische Zeitschrift* 39: 176–210 and 405–31.

Green, C. Application of theorem proving to problem solving. In *Proc. 1st IJCAI*, pp. 219–39, 1969.

H. P. Grice 'Logic and Conversation'. In P. Cole and J. Morgan (eds.), *Syntax and semantics, Vol. 3: Speech acts*, pp. 41–58, New York: Academic Press, 1975.

Grosz, B. J. and C. L. Sidner 'Plans for discourse'. In Cohen et al., pp. 416–44, 1990.

A. Heyting *Intuitionism: An introduction*. Amsterdam: North-Holland, 1956.

A. Heyting *Brouwer collected works (I)*. Amsterdam: North-Holland, 1975.

J. Hintikka *Knowledge and belief: An introduction to the two notions*. New York: Cornell University Press, 1962.

Hoffmann, J. and B. Nebel 'The FF planning system: Fast plan generation through heuristic search'. *Journal of Artificial Intelligence Research* 14: 253–302.

Konolige, K. *A deduction model of belief*. London: Pitman, 1986.

Kripke, S. 'Semantical considerations on modal logic'. In *Acta Philosophica Fennica* 16: 83–94.

Lewis, D. 'Scorekeeping in a language game'. *Journal of Philosophical Logic* 8: 339–59, 1979.

McCarthy, J. and P. J. Hayes 'Some philosophical problems from the standpoint of artificial intelligence'. *Machine Intelligence* 4:463–502, 1969.

Moore, R. C. 'A formal theory of knowledge and action'. In J. R. Hobbs and R. C. Moore (eds.), *Formal theories of the common sense world*, pp. 319–58, Norwood, NJ: Ablex, 1985.

Newell, A. and J. C. Shaw and H. A. Simon 'Empirical explorations with the logic theory machine'. In *Proceedings of the Western Joint Computer Conference*, 15: 218–239, 1957..

Newell, A. and H. A. Simon 'GPS, A program that simulates human thought'. In E. A. Feigenbaum and J. Feldman (eds.), *Computers and Thought*, pp. 279–93, New York: McGraw-Hill, 1963.

Nguyen, X. and S. Kambhampati 'Reviving partial order planning'. In *Proc. IJCAI*, pp. 459–66.

Pollack, M. E. 'Plans as complex mental attitudes'. In Cohen et al., pp. 77–103, 1990.

Ramsay, A. 'Speech act theory and epistemic planning'. In H. Bunt and W. Black (eds.), *Abduction, belief and context in dialogue: Studies in computational pragmatics*, pp. 293–310, Philadelphia, PA: John Benjamins, 2000.

Searle, J. R., 'What is a speech act?' In M. Black (ed.), *Philosophy in America*, pp. 221–39, Allen and Unwin, 1965.

Stalnaker, R., 'Pragmatics'. In *Proceedings of the Aristotelian Society* 86: 153–71.

Thomason, R. H. 'Accommodation, meaning, and implicature: Interdisciplinary foundations for pragmatics'. In Cohen et al., pp. 325–63, 1990.