MODELLING SPEECH ACTS IN CONVERSATIONAL DISCOURSE

Amanda Schiffrin

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy (Ph. D.)

The University of Leeds
School of Computing

May 2005

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

This copy is supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement

For my father, David Jorge Schiffrin,

Caminante, son tus huellas	*Traveller, your footprints*
El camino y nada más;	*Are the road and nothing else;*
Caminante, no hay camino,	*Traveller, there is no road,*
Se hace camino al andar.	*You make the road by walking.*
Al andar se hace camino	*By walking you make the road*
Y al volver la vista atrás	*And when you turn and look back*
Se ve la senda que nunca	*You see the path that can never*
Se ha de volver a pisar.	*Be trodden again.*
Caminante, no hay camino,	*Traveller, there is no road,*
Sino estelas en la mar…	*Only the wake that breaks the sea…*

Cantares, A. Machado & J. M. Serrat (translated by A. Schiffrin)

My mother, Margery Schiffrin,

World – how it walled about
    Life with disgrace
Till God's own smile came out:
    That was thy face.

Apparitions (Verse III), Robert Browning

And my brother, Roberto Schiffrin.

Happy the man, and happy he alone,
    He who can call today his own:
    He who, secure within, can say,
Tomorrow do thy worst, for I have lived today.
    Be fair or foul or rain or shine
The joys I have possessed, in spite of fate are mine.
Not Heaven itself upon the past has power,
But what has been, has been, and I have had my hour.

Happy The Man, John Dryden (translating Horace *Odes, Book III, XXIX*)

Abstract

Computational pragmatics and dialogue analysis is currently a rapidly growing area of research in computational linguistics. Over the past five years or so, initiatives in modelling pragmatic aspects of dialogue have led to considerably improved spoken language dialogue systems – so much so in fact that constrained human-computer interaction no longer seems out of the question.

One of the main drawbacks to such systems however is highlighted by the word 'constrained'. Human communication is seldom confined to answering questions or solving problems within a restricted field (such as train timetable enquiries, or route finding, for instance). How can one tell whether theories of dialogue that work well in domain specific, task-oriented dialogue, can be scaled up or expanded to deal with natural conversation?

In this dissertation I have carried out a critical survey of the various approaches to speech act modelling, detailing what I think are the strengths and weaknesses in the current theories. One very promising approach is that of using speech act analysis as a means of interpreting a speaker's intentions in producing an utterance. This then forms the basis for determining a hearer's response (following certain rules of conversational co-operation). I go on to present what is intended as a preliminary model, which is designed to capture the characteristic relationship and interaction of speech acts in conversational dialogue, especially those features which preceding research has failed to represent. Speech acts are defined by means of schemata that match the state of the prevailing conversational context space. Each possible context space is specified in the model for the performance of a particular speech act or acts; the representation of the context space is then updated accordingly.

I illustrate the theoretical model using real conversation, collected during the course of this research, and compare its performance against the analysis of a 'benchmark' conversation, highlighting where the model falls short and how it could be improved in the future. I will argue that the model provides a powerful formalism for the characterisation of a wide variety of different basic speech acts.

Declarations

Some parts of the work presented in this dissertation have appeared in the following conference and workshop publications:

Schiffrin, A. and Souter, D. C. 2002. 'Evaluating a Pragmatic Model of Speech Act Assignment Using Spoken Language Corpora', *Paper presented at the 23rd International Conference on English Language Research on Computerized Corpora of Modern and Medieval English (ICAME)*, Göteborg, Sweden 22-26 May 2002.

Schiffrin, A. and Millican, P. J. R. 1998. 'Identifying Speech Acts in Context', *Proceedings of the 2nd International Workshop on Human-Computer Conversation*, Bellagio, Italy, 13-15 July 1998.

Research presented in this volume extends and supersedes earlier work carried out on the Commitment Slate Model reported in the following:

Schiffrin, A. 1995. *A Computational Treatment of Conversational Contexts in the Theory of Speech Acts*. Leeds: The University of Leeds (M.Sc. Dissertation).

There are two further points that I would like to make clear from the outset.

The first is that I have used British English spelling as consistently as possible throughout this work. Cases of American English spelling should only be found in quotations and references.

The second is that I have made use of the generic third person singular throughout this dissertation. This may not be very politically correct, but it is linguistically correct, and besides sounds more natural, which is, in my view, of greater importance. To those for whom this is an issue please read 'he/she' instead of 'he', 'him/her' instead of 'him' and so on.

Contents

List of Figures

List of Tables

Acknowledgements

Over the years I have accumulated a variety of debts of gratitude, which I would like to acknowledge here. Primarily I would like to thank the following people:

- My supervisor, Clive Souter, for his constant and comprehensive help and advice, even under difficult circumstances, and for guiding me when I was losing my way. I would like to thank him for being more than my supervisor, for being a good friend.

- The School of Computing (staff and postgraduates), for providing a congenial environment in which to work, and for employing me to support my final years of study.

- The Engineering and Physical Sciences Research Council, for funding my research for three years, and for allowing me to expand my horizons by placing my grant in abeyance while I spent a year working in the Natural Interactive Systems Laboratory, University of Southern Denmark, Odense.

- Peter Millican, for talking me into studying computing in the first place back in 1990 – without his primary encouragement and help I would never have got this far. Also thanks go to him for aiding in the development of the ideas in my Masters degree, which sparked off the themes of my doctoral research.

- My proofreaders: Clive Souter, Katja Markert, Roberto Schiffrin, Margery Schiffrin, Jackie Knight and Keith Hobley. Also Catherine Greenhill, for typing up some of my notes for me.

- Jackie Knight in particular for her unwavering love, support; for sitting with me and keeping me company through the writing up process; and for numberless dinners.

- All my other friends for their understanding and care.

Finally, I would like to dedicate this work, with much love, to my family: to my parents, David and Margery Schiffrin who have given me so much – there is no way to thank them adequately for all they have done for me over the years; and to my brother, Bob, for being himself and for being there throughout.

Chapter 1

Introduction

> I know you believe you understand what you think I said, but I am not sure you realise that what you heard is not what I meant. (Professor Thomas L. Martin)

I begin my dissertation with this tongue-in-cheek quotation because it begs the question that lies at the heart of my research: namely, how do we understand what we mean by what we say?

On the face of it, the act of communication – such a commonplace phenomenon in the experience of most human beings – seems simple. One talks, someone listens, and then replies in turn with some relevant contribution, and so on. We do this as naturally as any other of our acquired cognitive skills and abilities in life. The man in the street might then be forgiven for assuming that the underlying structures that govern the performance of such an everyday task would also be correspondingly simple, and as such, readily amenable to logical and computational expression and emulation.

To those in the habit of studying language for a living, however, it comes as no surprise or shock to discover that the underlying mechanisms, for understanding what we intend to convey to each other in speaking, are extremely complex indeed. So much so that, 40 years on from the publishing of Austin's (1962) seminal work describing his theories about "utterance as action", there is still heated debate even about which approach is the most appropriate for this task.

1.1 The Problem with Language

The problem of how to engage in effective communication is not immediately obvious. In children, the development of language and the acquisition of the rules of language are just as important as the development of other cognitive processes and social behaviours. According to some psychologists (Piaget 1955 and Vygotskii 1962, for example), language is an integral part of cognitive development, along with sensor and motor processes, pre-conceptual thought, symbolic thought, intuitive thought, etc. It is argued that language is so closely connected to our mental development that it is in fact the very framework and building blocks of our intelligence and reasoning abilities. There is some discussion as to the extent of the co-dependency between language and thought, but there is little doubt that it exists, especially in the acquisition of the more abstract levels of thinking; language seems to be the medium through which intelligence and mental development occur and by which thinking may be gauged (Aitchison 1989).

With this in mind, it would seem of considerable importance for those in the field of artificial intelligence (AI), whose aim is to replicate human intelligence and behaviour by computational means, to be able to give some account of our facility and preference for the use of spoken language as our primary means of communication. Surely one of the preconditions for a computer to be deemed intelligent is that it should be able to 'understand' natural language (where 'understand' may be interpreted as 'to be able to process by means of rules into a form which is readily accessible to the computer'). Needless to say, this is no simple task. Language is the 'verbalisation of thought' and is neither straightforward to process nor to break down into manageable pieces. Humans themselves spend decades and many thousands of conversations to fully acquire all the intricacies of any given language. It may be that any system for computer language acquisition would require the same range of data input (although obviously computers are capable of much faster processing than human beings) for a similar learning process to take place.

Language is not merely a collection of words to be processed and acted upon according to set rules. Quite apart from all the syntactic, grammatical and semantic difficulties encountered when trying to work out the structure of language, there are many other extraneous influences that govern any sentence; such as intonation, the character of the speaker, the occasion, the motives of the speaker, the use of jokes, sarcasm, metaphor, idioms and hyperbole (when words may not be taken at face value), etc. The understanding of language necessarily entails some understanding not only of the world, but also of the conversational 'context' of an utterance. This point is made by Reichman (1985), who says, "in addition to our knowledge of sentential structure, we have a knowledge of other standard formats (i.e. contexts) in which information is conveyed."

How could such information be represented computationally? Certainly with the current computing technology available, any explicit representation of world knowledge and context would take far too much time to consult and be too exhaustive to be efficient; neither would it be a true model of human thought, in which a great deal of deduction and inference takes place. Language itself is riddled with shortcuts and contraction, devices that aid in the rapid conveyance and processing of ideas. To enable a computer to work similarly, knowledge would have to be structured in such a way as to reflect the relations between words performing different functions in language and to allow speakers the possibility of constantly updating the information stored.

1.2 Conversational Coherence

In order to get a clearer idea of the issues involved before further addressing these observations, it is necessary to consider the way humans keep track of conversation. Socio-linguists have put forward the idea (originally conceived by philosophers of language – see Chapter 3) that all speech can be seen as a variety of 'social action', such as greeting, promise or declaration, etc. Labov (1970) says that there are "rules of interpretation which relate what is said to what is done" and it is upon the presupposition of these rules that any given dialogue can be considered coherent or incoherent. He gives an example of incoherent interaction between a doctor and a schizophrenic patient (from Laffal 1965):

Doctor: What's your name?

Patient: Well let's say you might have thought you had something from before, but you haven't got it any more.

Doctor: I'm going to call you Dean.

Labov contrasts this example with one that shows only an apparent lack of coherence:

A: What time is it?

B: Well, the postman's been already.

In this instance, it is assumed that **B** is <u>not</u> schizophrenic and is making some attempt at answering **A**'s question. **A** must try to infer **B**'s meaning. The first assumption may be that **B** has no direct means of knowing precisely what the time is (e.g. **B** is not wearing a watch today) and therefore cannot give an exact answer. However, the postman always arrives at around 11am, and **A** knows this fact. So, as the postman is declared to have already come, **A** can deduce that it is some time after 11am. This information at least narrows down the possibilities, even if it is not maximally informative. So with this assumption, **B**'s answer can be said to <u>count</u> as a declaration of the fact that it is past 11am and is not just an arbitrary statement about the way the world is.

This is not the only interpretation of the above. **A** might be anxiously waiting for the post to arrive for some particular reason (e.g. exam results); if **B** knew this to be the case, his answer can be taken at face value. However, the underlying content of **A**'s original question would then be 'What time is it because I want to know whether the postman has come?'.

Further evidence for "utterance as action" may be found in the use of words such as 'because' as conjunctions between questions and propositions:

'What's the time because I'm waiting for the postman?'

Here 'because' is not used as a propositional connective (P because Q) or as a linking conjunction between two clauses, but is used to give a reason or explanation for asking a question. So our understanding of the question is based "on our assumption that a reason is being expressed for an action performed in speaking" (Brown and Yule, 1983). Both the action and the reason for it are made known to the speakers by their location within a conventional structure of spoken interaction.

It is thus that two apparently unconnected sentences, lacking in cohesion from the point of view of explicit indicators, may still be interpreted as a coherent piece of conversation:

A: That's the telephone.
B: I'm in the bath.
A: OK.

Widdowson (1978) argues that it is only by analysing each part of the dialogue and extracting the action which each is performing within the dialogue, that it is possible to accept this conversation as coherent. So we analyse the conversation in the following way:

A requests **B** to perform action (answer the phone).
B states reason why he cannot comply with request.
A undertakes to perform action (or at least accepts non-compliance).[1]

Here we begin to perceive how conversations can be analysed as a series of transactions and conversational moves within a contextual structure. So far, "utterance as action" has been described in a very informal way in order to introduce some of the ideas. I shall now consider the concept of a speech act in slightly more detail.

It is important at this stage to look at an example of the phenomena I wish to deal with, and to examine briefly why it is of such importance to this dissertation, and to natural language understanding as a whole. I shall cover the philosophical foundations of speech act theory in detail in Chapter 3.

---

[1] We might instinctively feel uncomfortable saying with certainty that this is the correct interpretation of this example conversation. This is a rather simplistic analysis and tells us nothing about the circumstances under which such an utterance, made in the declarative form, can be taken as a request. But we shall come back to this point later, in Chapter 4.

1.3 Utterance as Action – Speech Acts

As we have already noted, in order to interpret any natural language utterance within a normal human conversation, it is not enough to know the grammatical category of all the words in the utterance, nor the conventional meaning associated with each word, nor even how such meanings combine to form an overall sentential meaning. Before being able to ascribe a particular meaning (specifically that intended by the speaker) to any utterance, a hearer must have a clear idea of the context in which such an utterance occurs. For example, if I say:

'It is cold in here.'

I could be intending my audience to understand any number of meanings. I might simply be performing an act of informing or asserting. Or I might be indirectly asking one of the hearers to shut the door, or bring me a coat. I might be intending sarcasm if in fact it is very warm, or changing the subject at the entrance of another person into the room, etc. One can conceive of many interpretations for even such a seemingly straightforward utterance as the example given above. The linguistic term for the intention or force of an utterance is its **speech act** (also known variously as 'illocutionary act', 'dialogue act', 'discourse act' or 'speech function' according to preference[2]). The problem then is how to formalise this idea of dialogue context in such a way that a computer program could recognise the speech act being performed by an utterance in a conversation and so begin to grasp some of the subtleties of human interaction. It is this problem that I am primarily concerned with elucidating here.

1.3.1 Why is it Important to Recognise Speech Acts in Discourse?

Speech acts are the underlying actions we perform when we speak[3]. Some examples are: INFORM, COMMAND, PROMISE, REFUSE, etc. Recognising the speech act that is being performed in the production of an utterance is important because it is the speech act that to some extent tells us what the speaker intends us to do with the propositional content of what he says.

The identification of the speech act that is intended by the production of an utterance is vital then as it provides appropriateness constraints for our responses. By this I mean that after every utterance, conversational expectations are created (either implicitly or explicitly) which serve us in understanding later conversation, in producing a relevant and appropriate response, and, very importantly, in being able to identify when and where a conversation goes wrong.

---

2 Throughout this dissertation I shall be using the term 'speech act' to subsume all other terminology referring to the same phenomena, except when explicitly stated otherwise.

3 In later chapters, I will cut down and refine the definition of a speech act.

Not only this, but if we cannot understand the function intended by the production of a certain utterance, then we will also be unable to form opinions about the position of a speaker with respect to the content of his utterance. So, recognising speech acts could be essential for ascribing the correct[4] beliefs and goals to a participant, for gleaning background knowledge of that participant and thus for being able to build on the knowledge gained from the current conversation in order to facilitate future interactions with that speaker.

In essence, it is as if we build a map of the discourse (and the information managed therein) as we speak. It is the structure of this 'map' that is the subject of this dissertation. For I will argue that without taking this feature of language into consideration, it will not be possible to teach computers to mimic human conversational behaviour in all its complexity.

1.3.2 How Do People Recognise Speech Acts in Discourse?

We can generally[5] infer the speech act from the following three properties:

(1) The *content* of the utterance: That is to say the proposition expressed by the utterance. Defining what that is is not always easy. If I say "Shut the door" I mean "Bring it about that, by some (future) action, the proposition 'The door is shut' becomes true in the context of the physical world". Of course we also use background knowledge and knowledge of the speaker to deduce the content of an utterance. I assume for simplicity's sake that this is possible to do – although not necessarily unambiguously so.

(2) The *force* or *mood* of the utterance: This is achieved by what I will call *descriptive*, *prescriptive* and *requestive* markers, which roughly correspond to the traditional mood types *declarative*, *imperative* and *interrogative*.

(3) The *position* of the utterance within a conversation.

It is this last that is of most interest to me. The placing of an utterance in a conversation is important because the same content can be interpreted as different speech acts depending upon its position relative to other utterances in the same context. This inevitably affects the understanding of the function of the utterance. For example, take the following hypothetical conversation:

---

[4] In Chapter 2 I shall discuss whether it is ever possible to be sure that one has indeed correctly identified the speaker's underlying thoughts, beliefs and motivations, but this is not important here.

[5] These are not the only ways we do it – we also use other linguistic markers, such as inflection/intonation, cue phrases, use of modals, etc. These are not important here.

(1) **A**: The door is shut. (ASSERT)

(2) **B**: The door is not shut. (DISAGREE)

(3a) **A**: The door is not shut. (CONCEDE)

(3b) **A**: The door is shut. (INSIST)

This is a very crude example of how exactly the same utterance, even uttered in the space of a couple of turns, must be interpreted as performing quite different functions in the conversation. Obviously no-one actually ever speaks with the content fully specified in this way, this is just intended to be representative; if this conversation were to take place, it would be realised in a much more natural manner. However, it is clear that the only way one can attribute different speech act meanings to the same utterance (or the same propositional content of an utterance) is by reference to the conversational context and to the order in which it comes in the dialogue. For example if on the performance of (3a) we only consider the preceding utterance, then the act would be deemed one of agreement rather than concession. Without some method of retaining where we are in the dialogue, we would be unable to interpret the current speech act satisfactorily.

Sometimes people are unable to identify the intended speech act. If we are unsure of the role an utterance is playing within a dialogue, we will need some form of clarification from our co-participants. Below is a brief excerpt from a real conversation, taken from the spoken section (KB0) of the British National Corpus (BNC) showing this behaviour:

A: You enjoyed yourself in America.

B: Eh?

A: Did you?

B: Oh, I covered a nice trip, yes…

Here we can see that **A** has asked an indirect question by making a statement about **B**'s attitude towards his trip to America (that he enjoyed it), a statement which **A** is clearly unqualified to make (not being able to mind-read). **B** signals his incomprehension of the function of **A**'s statement. **A** then correctly judges that this is a request for clarification (rather than a signal that **B** did not hear **A**'s original utterance), and clarifies it with an explicit question form 'Did you?'. This is presumably a contracted form of 'Did you enjoy yourself in America?'. **A** can omit the content here because that is already in the conversational context; although, note that the elliptical form is actually the ungrammatical 'Did you enjoyed yourself in America?'. This however poses no problem to **B** who makes the correct inference and now feels confident enough of the function of the utterance to be able to answer the question.

In actual fact this might just be a ploy by **B** to gain time to think before answering **A**'s question, but I include this example here because, although I do not think this kind of indirectness would

trouble many people, I wish to show that human beings are sometimes unsure of the correct interpretation of the speech act and have to recover in real time. I believe that, in the body of theoretical work on speech act modelling at least, too great an emphasis has been placed on determining the exact, correct act being performed, and on producing an utterance that has one correct interpretation. The evidence is that 'real' conversation is robust and resilliant enough to cope with such problems on the fly.

June: Shut door.

Albert: I can't I'm going to er…

…<Gap of a few minutes' unrelated conversation>…

Albert: Do you want tea June?

June: No I'd rather have coffee.
I mean I'd rather have tea, sorry, yeah.

Albert: I wished you'd make up your mind.

June: Shut the door.

Albert: I can't shut door. <*pause*>

June: I will.

Albert: Which teapot is it? <*pause*>

June: Does it matter what teapot you make it in? <*pause*>
In the metal one.

…<Gap of a few minutes' unrelated conversation>…

Albert: I've shut the door now. I've finished running about.

Conversation 1.1 Snippet of conversation between June and Albert.

It is all very well talking about this in the abstract, or with very simple examples, but if one looks at any sample of real conversation, one gets an idea of the scale of the problems of deciphering the function of utterances in naturally occurring language. In order to illustrate some of the obstacles facing function recognition more clearly, I include here a brief example conversation between a husband and wife taken from the spoken section (KB1) of the BNC (see Conversation 1.1 above). I have here cut out two sections of conversation about other things to try and keep the different strands of conversation manageable. Even so, what are we to make of this? I have listed some of the features of interest below.

Nested threads of conversation: We have (ignoring the excluded conversation) two main related threads of interaction: (a) the thread about shutting the door, and (b) the thread about making a cup of tea. These threads are not dealt with neatly, one at a time, but are maintained open at some level of activity throughout the conversation. The different topics are nested within each other, so that it is a wonder that the participants are able to switch their focus of attention so readily between the different threads and manage to keep track of where they are up to.

Self-contradictory utterances: June displays some strange self-editing behaviour in her second utterance given above, in which she manages in one turn to reject Albert's offer of a cup of tea, express a desire for coffee instead, change her mind (signalled by the cue words 'I mean'), express a desire for tea instead (of the coffee), apologise (for the confusion?) and finally accept Albert's original offer. This (seemingly schizophrenic) behaviour is primarily a result of the immediacy of spoken language. June simply has not had enough time to think through her answer before talking, so she does it while talking, with the consequent effects seen here. In another conversational setting this kind of informality might not be appropriate, but in a home situation such an utterance is perfectly normal, despite Albert's grumbling about June's inability to make up her mind. Note this is interesting in itself, as it shows that Albert is able to recognise the underlying structure in June's rather incoherent turn. In fact he picks on her for not being able to answer him simply and immediately, despite the fact that she does eventually respond to his offer.

Irrelevance and lack of co-operation: We can also see in this conversation clear examples of what looks like conversational irrelevance, and, in fact, downright lack of co-operation in some cases. The interleaving threads of conversation seem on the surface to lead to irrelevance. The repeated refusal to comply with the request to shut the door (albeit issued rather brusquely as a command), and June's rather cross-sounding 'Does it matter what teapot you make it in?', all show an apparent lack of co-operation.

From a sociological point of view, we might find it interesting to note that the kind of behaviour I have discussed in this conversation, in fact rather typifies the interaction between a husband and wife, especially of long-standing relationship. What is it about this conversation (which let us face it, is just a collection of written words on a page) that would almost certainly suggest to an adult casual observer the sort of relationship that exists between these two people? This is a little beyond the scope of my research, but is worth noticing here because it exemplifies how language is structured around patterns of behaviour, and this in itself helps us interpret each other's meaning in conversation.

So, the question remains, how do we manage to co-ordinate our conversation at all? If we break down this example conversation further, we will see that there is a high level of organised structure present which allows us to track what is going on (see Figure 1.1). Without going into too much detail at this stage, what seems to happen is that both participants in the conversation maintain the various threads in talk, keeping a mental placeholder when the current thread is deemed unresolved, updating the context accordingly as each utterance (and co-ordinated action) is made, until each thread is finally resolved and closed, remains unresolvable (in the case of disagreements for instance), or is just simply forgotten.

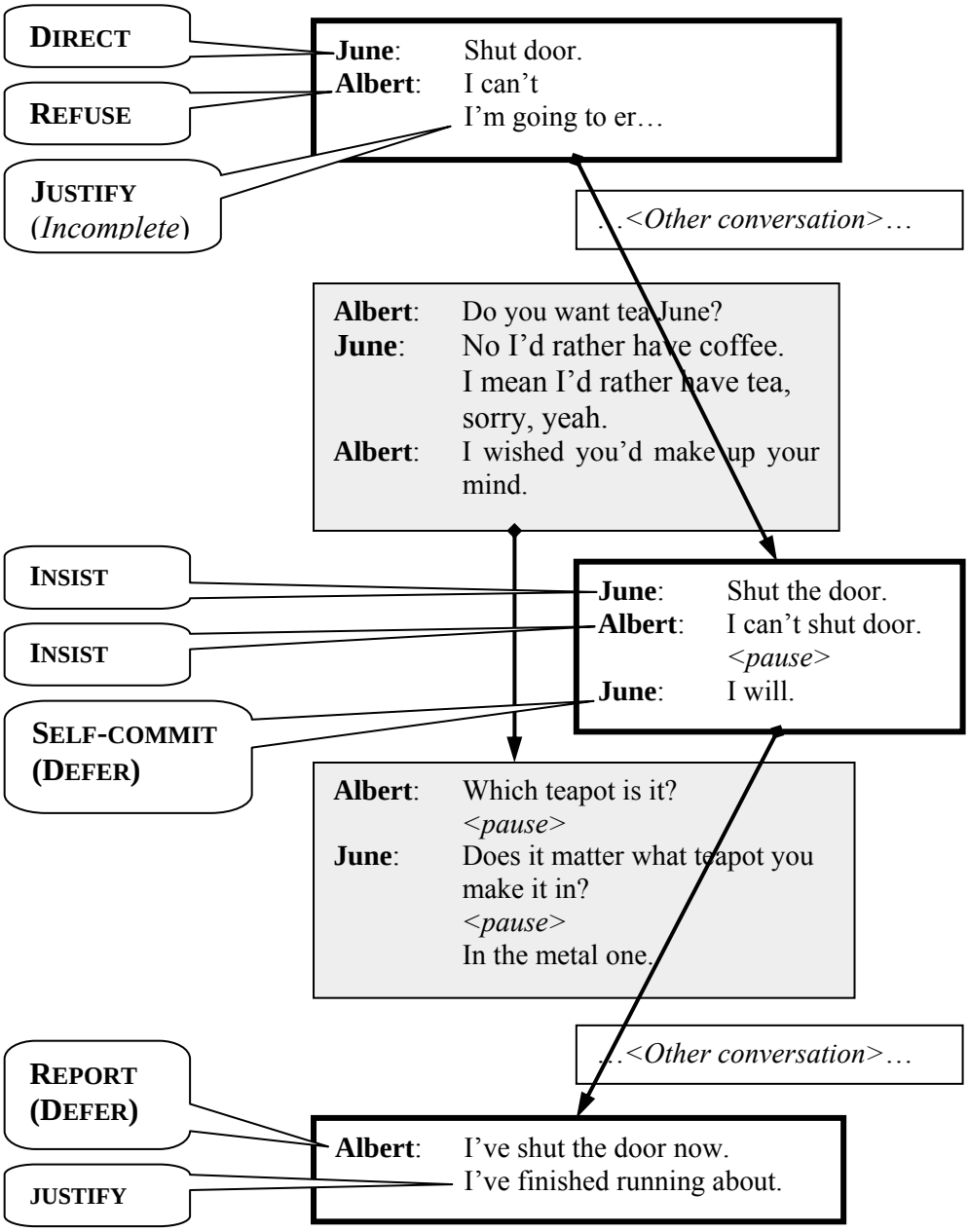


Figure 1.1 Breakdown of Conversation 1.1.

I hope to show that what seems at first to be rather random, disconnected and even uncooperative conversational behaviour, in actual fact exhibits a high degree of organisation and collaboration. An example of this is when at the end of the conversation given here, Albert does indeed comply with June's request to shut the door, even though June has since committed herself to carrying out this action. This is because humans spend a lot of their time problem solving, both for themselves and co-operatively for other people. Whenever they are able to do so without prejudice to their own goals, people are apt to adopt another's goals themselves (obviously within reason, and as long as those goals are within socially acceptable bounds, etc.). Again I am straying from the main line of enquiry of this dissertation here, although such observations are relevant when considering the interpretation of utterances in conversation.

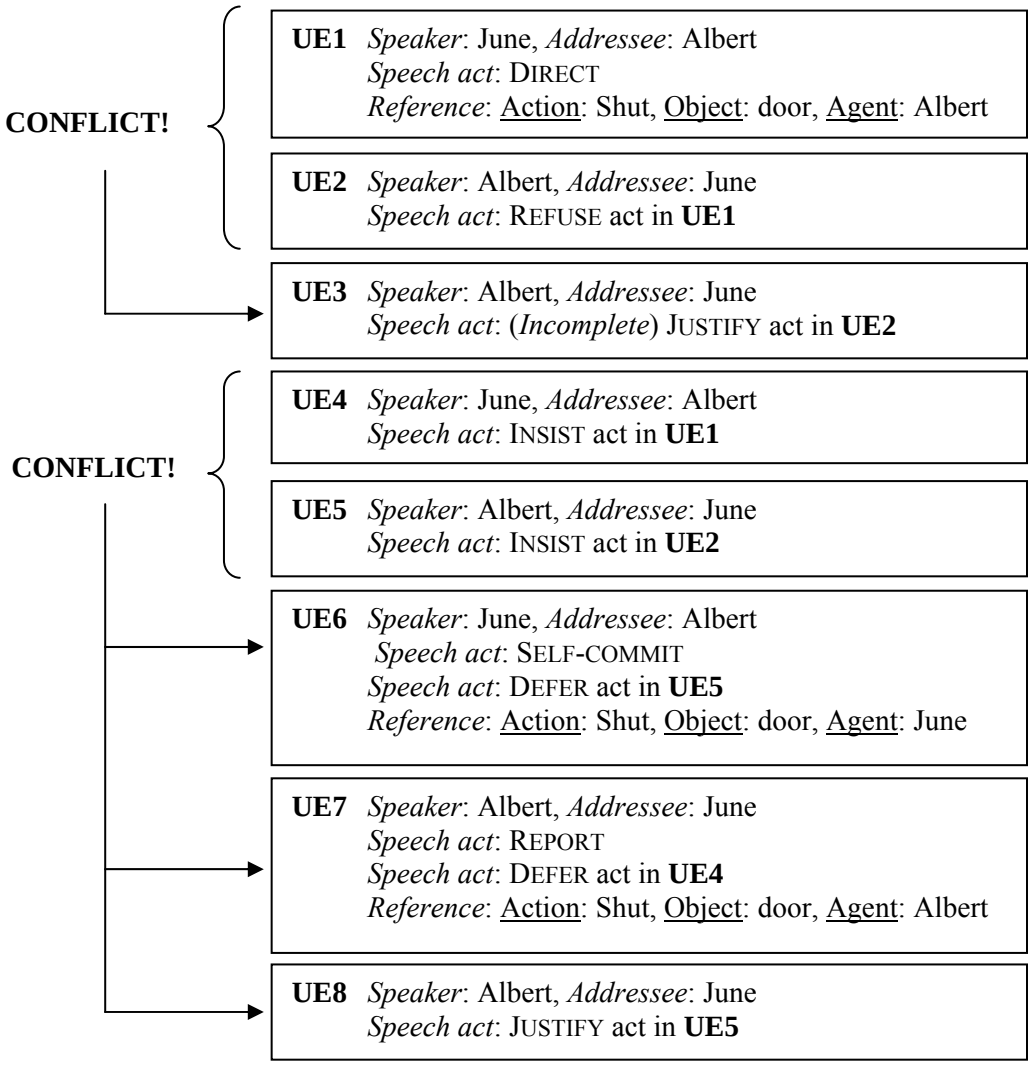


Figure 1.2 Analysis of thread in Conversation 1.1[6].

If we then look at one particular thread of the conversation concerning the shutting of a door (highlighted in bold borders in Figure 1.1), we can see that the correct interpretation of any one of the utterances crucially depends on the rest of the conversation. This is demonstrated graphically in Figure 1.2.

I should note here that, having not taken part in this conversation, nor had access to the recording from which this transcription was made, I am not fully aware of the context and background of the talk. I cannot be sure therefore that I have not misunderstood what is really going on. Even the tone of the conversation is lost. This couple might be constantly sniping at each other as evinced by various retorts made (like 'I wished you'd make up your mind' and 'Does it matter which teapot you make it in?'), and by the lack of a lot of commonly expected

---

6 **UE** here stands for 'utterance event'. I use the term 'event' because every utterance is firmly anchored in time. The sequence of events is crucial for determining the act being performed, as I discuss in the following section.

politeness forms in their language. Or the retorts could just be seen as banter, and the apparent discourtesy could just mean that they are so comfortable with each other that they feel they can omit certain norms of politeness.

With this in mind, any of the analysis and assumptions that I have made about this conversation can only ever be conjecture. This does not in any way invalidate the issues I have raised in the discussion above (for I have created my own context for this conversation in which the observations I have made are applicable), but it does suggest that there is no way of finding out the 'true' intentions of the speakers here for having used these utterances. This issue will be of consequence in later considerations about what data is appropriate for studying these types of conversations (see Chapter 7).

Having described in some length here some of the problems that interpreting conversation poses, I now turn to assess what the main goals of this work should be in light of the discussion above. I shall focus mainly on the qualities of my research that are not adequately covered (in my view) in other related work in this area.

1.4 Aims and Objectives

In this dissertation I hope to provide a new approach to speech act recognition, built on the foundations of existing work, and in particular extending earlier research as detailed in Schiffrin (1995). I shall not here be concerned with making quantitive claims about my observations. My chief aim will be to point out patterns of behaviour that are repeatedly to be found in the data, and in so doing to discover something new about the nature of conversational behaviour as shown by the interplay of speech acts. I shall not be looking at the detailed distribution of the occurrence of these behaviours.

I intend to try to give an account of the following elements of speech act theory that have hitherto not been satisfactorily dealt with, and also to accommodate these elements within a generalised approach to identifying speech acts in context.

(1) Sequences of speech acts can have dependencies of more than just one conversational turn. Unlike the theory expounded by conversational analysts concerning two-sequence pairs called 'adjacency pairs' (see the discussion in Chapter 5 for further details), I argue that the recognition of speech function relies possibly not only on more than one previous act in the conversational context, but also requires an account of the order of performance of these prior acts. Although conversational analysts do account for nested adjacency pairs (such as questions whose answers depend on the answer of sub-questions, etc.) and other such sequences, there is no indication of how speech acts alter the context and affect the interpretation of following speech acts, except with reference to the expectations that

certain acts throw up. I argue that this is insufficient to explain the complex relationships to be found between speech acts in naturally occurring conversation. Speech acts are sequential, co-dependent events, which do not necessarily occur in adjacent positions (one only has to refer back to the conversation in Section 1.3 to see examples of this).

(2) I make a distinction between the underlying act, and the recognition of the *behaviour* that that act represents or counts as within the conversation. In fact, I will suggest that the term 'speech act' has been used to cover a wide and varied spectrum of phenomena – so much so that it is difficult to say what the definitive, or archetypal 'speech act' really is. As I argued in Schiffrin (1995), different levels of theoretical abstraction have been conflated and described by one overarching label 'speech act'. I shall be discussing this in Chapters 6 and 7 when I look at other speech act schemes that have been developed.

(3) I shall apply and validate my observations by looking at real speech as it occurs in conversations. There are difficulties and obstacles consequent to this decision, however, I am in agreement with Leech here, who says, "any account of meaning in language must… be faithful to the facts as we observe them" (1983: 7). I see little point in developing a theory that appears intuitively to work, but does not actually fit the data. All too often theoreticians working in this field can have this criticism levelled at them: that, in using made up, subjectively informed examples, they try to make the data fit the theory. It is my opinion that if an idea about how conversation is constructed fails to match the observed reality, then it is not of much use. While I am not claiming that there may not be grains of truth to be found in such theories, I am claiming that their position is weakened without reference to actual evidence taken from real dialogue between people. I am conscious in stating this that I am placing myself in the odd position of intermediate between two extremes, as represented by those studying language with only their own intuitions to guide them, and those who only look at the statistically relevant distributions of specific linguistic elements within language. I hope in this research to occupy the uncomfortable middle ground between these two approaches, which are typified (though not wholly so) by generative and corpus-based linguistics.

(4) I hope to show that, just as the form of a sentence is governed by a grammar, at a higher level, discourse too can be said to be composed of units that work together.

(5) Often in theories of discourse, the role of the hearer of an utterance is overlooked or ignored. Clark (1996) has proposed a theory of joint action to take into account the hearer's role of *consideration* (as well as the speaker's role of *proposition*). While I agree with this approach in principle, I think one must analyse a hearer's perspective from a single, subjective point of view. After all, we are only ever able to assess a hearer's understanding from our own model of interpretation of utterances. This observation

reflects the nature of our constantly switching roles between that of speaker and hearer in conversation. I shall be arguing that we can never 'mutually' know each other's intentions in speaking.

(6) Many speech act analyses are based on the fulfilment of certain *felicity conditions* that must obtain before the correct (or 'happy') performance of a speech act can be deemed to have taken place. Some of these felicity conditions refer to the internal states and beliefs of a person, and the sincerity of their utterance in relation to these. These types of felicity conditions I shall attempt to reformulate in terms of conflict and contradiction, in order to be able to do away with the necessity of an explicit model of belief for the ascription of certain types of speech act.

1.5 Scope of Research

There have been a number of other attempts to try to define and recognise contextually the speech acts being performed within a dialogue (I shall be discussing these in detail in Chapter 6 and 7). The success of these different approaches has been varied. Most have been characterised by being restricted to a particular domain or specialised task type. There has also been an emphasis on the manipulation of the beliefs and goals of the participants within the dialogue in order to interpret the speech act under execution. Whilst I do not wish to undervalue the importance or impact of taking beliefs into account for any theory of contextual representation, I believe that often a crucial step has been overlooked or missed out altogether. Regularly in the ascription of speech acts to utterances, one need not appeal to the resolution of complicated inference structures at all, but, as I hope to show in this dissertation, they can much more simply be accounted for by the constraints and expectations invoked by the immediate conversational context. Essentially I hope to place this approach on a more principled footing.

I have tried to keep the field of inquiry as open as possible, but it would be fitting here to delimit the kind of data I have looked at. Some of the features that characterise my approach to analysing conversation are listed under the following headings (this list, is more fully defined and discussed in Chapter 7). I shall try to justify the choices I have made throughout this dissertation.

(1) Generic (as opposed to specific) context of situation.
(2) Open, general (as opposed to domain restricted).
(3) Non-goal-driven (as opposed to task-oriented conversation).
(4) Theoretically structured speech act annotation scheme, backed by observation.
(5) Speaker/hearer distinctions, but from a role-change perspective.
(6) Acoustically informed annotation process (backed wherever possible by first hand knowledge of the participants and situation).

(7) Accommodation of any number of speakers within my model[7].

(8) Dealing with long distance dependencies and discontinuities in the data.

(9) Distinguishing between different levels of abstraction.

(10) Semantic/pragmatic distinction.

(11) Syntactic/pragmatic distinction.

Having set the scene for the coming discussions, I now wish to look briefly at the scope of the subject area as a whole in order to position my own work within it.

1.6 The Need for a Unified Theory of Discourse Interpretation

The importance of conversational context for the purpose of recognising a speaker's speech act (and in so doing, identifying the intention or function of an utterance within a conversation) is widely acknowledged in all fields related to the understanding of natural language. Over the past thirty years or so there has been considerable discussion concerning the formal representation of context, and the role it may play in interpretation. The disciplines that make claims to the study of discourse are extremely cosmopolitan. The interdisciplinary nature of this research has meant a need to read very widely in the literature of a variety of sub-fields of computational linguistics, the philosophy of language, sociology, ethnography, discourse-based linguistics and cognitive psychology.

Speech act theory has been applied to a variety of different computational architectures alone, some examples of which include:

- A proposed computer programming language called Elephant 2000 (McCarthy 1990), "for writing and verifying programs that interact with people (e.g. transaction processing) or interacting with programs belonging to other organisations (e.g. electronic data interchange)".

- KQML (**K**nowledge **Q**uery and **M**anipulation **L**anguage) an agent communication language and protocol for exchanging information and knowledge (Finin et al. 1997). It is used to signal agent attitudes about information such as querying, believing, stating, requiring, subscribing, etc. KQML aids in the dissemination of knowledge at run-time between agents by providing both the format and the protocol for handling messages.

---

[7] Note that this is only a theoretical accommodation of any number of speakers as in practice conversation becomes extremely complex with more than three participants (due in part to conversational splits).

- Other negotiation and argumentation protocols for co-operative and resource-sharing systems in autonomous agent theory (Esa and Lyytinen 1996, Parsons and Jennings 1996, Sierra et al. 1997), as well as other information systems theories (Chang and Woo 1994).

- In machine translation (Alexandersson et al. 1998, Tanaka and Yokoo 1999) where speech acts are used as a translation invariant.

- Dialogue management systems (Ardissono et al. 1998, 2000, Hagen and Popowich 2000, for example).

- Belief modelling: for the purposes of recognising speakers' plans and goals in natural language processing (Cohen and Perrault 1979, Allen 1983); for planning coherent natural language generation (Appelt 1985); for the ascription of mental attitudes (Lee and Wilks 1996); for the identification and resolution of conflicting belief structures (Lee and Wilks 1997); and many more (see Chapter 6 for a more thorough overview).

These very different applications of speech act theory have met with varying degrees of success, often reflecting the nature of the tasks for which the theory has been used. There are two main criticisms one might level at current models of speech act theory. First, that they are in general very domain-dependent, aimed at solving a specific problem (and in most cases limited in scope for this reason); and secondly that they often involve complex belief structures which lead to combinatoric complications when attempting to implement any non-trivial reasoning automaton.

Each of these approaches highlights different aspects of the theory, whilst very few have made an effort to take a look at the problem as a whole – the problem of how to communicate effectively. Most approaches bring their own theoretical biases to bear upon the puzzle of recognising speech acts. Although I am wary of claiming that this is not the case here, I will say that I have attempted to maintain (appropriately) a pragmatic stance.

The model I present in this dissertation aims to provide a skeleton on which a more substantial and fully formed theory of speech acts might later hang. It aims to avoid both the problems of particularity and complexity mentioned above, by providing a simple structure that represents only the relatively formal features of speech acts in general, and thus abstracts from the complex mental realities that in any concrete instance might underlie their use. In the next chapter I shall begin to outline some of the basic assumptions that I take for granted in the rest of this dissertation about the nature of human communication.

Chapter 2

What is Communication?

> COMMUNICATE *v*.
> [f. L. *commūnicāt-* ppl. Stem of *commūnicā-re* to make common to many, share, impart, divide, f. *commūn-is* common + *-ic-* formative of factitive verbs. The earlier Eng. spelling partly followed the variants of COMMON *a*.] *Oxford English Dictionary* (Online, 2002)

As we can see, the etymology of the word 'communication' comes from the Latin 'to make common to many'. This has been the main assumption behind a considerable number of linguistic theories – that in some way in speaking we share knowledge with each other in a common pool. In fact this idea survives (and is still subscribed to in varying degrees) up to the present day. I believe that this has also been the root of a misconception about communication. In this chapter I wish to sow a few doubts about some established ideas regarding what is communication, and by implication therefore also cast a sceptical glance at the resurgence of related ideas (such as 'mutual belief') in the discourse analysis communities.

2.1 Is 'Perfect' Communication Possible?

One of the earlier philosophers of language, John Locke, in his *Essay concerning human understanding* ([1689] 1971) expresses the idea of the commonality of language understanding in its most uncompromising form.

> Unless a man's words excite the same ideas in the hearer which he makes them stand for in speaking he does not speak intelligibly. ([1689] 1971: 262)

Interpreting this in its strongest form would assert that communication is in essence a means of thought transfer – a speaker encodes his thoughts into words, transmits them through the sound-waves of speech, and the hearer decodes the information and thus gains a replica of the speaker's original thoughts.

Locke himself was aware that achieving such flawless, ideal communication was highly implausible, subject to failure and other difficulties. However, this view was widely followed by Locke's contemporaries and still survived in various guises until the beginning of the last century. I would suggest that echoes of this way of viewing communication are still found in many modern day theories. Locke pointed out that, "since different individuals had different experiences, they used and understood words in different ways" ([1689] 1971: 300). He also realised that, as our ideas increase in complexity and abstraction, it becomes harder for the intended meaning of a speaker's utterance (which might not even be completely clear to the

speaker himself) to be deciphered in precisely the same way by the hearer. Ziff (1969: 233) says that natural language "does not ever have, not even at an arbitrary moment of time, a static fixed store of word senses". The meanings of a particular word are constantly expanded.

However, seemingly satisfactory communication can and does take place without the need for perfect understanding (as is the case with children for example). Quine suggests that actually all language is by nature vague: "Vagueness is a natural consequence of the basic mechanism of word learning" (Quine 1960: 125) and physical objects "will be vague in two ways: as to the several boundaries of all its objects and as to the inclusion or exclusion of marginal objects" (ibid.: 126).

Often even adults cannot distinguish the exact sense of certain words. For example, you might ask someone to distinguish instances of different types of flowers (by asking them to pick them out in a garden): rose, chrysanthemum, hyacinth, aster, hydrangea, lupin, etc., and they may be embarrassed to find out that they are unable to do it. Yet, in a conversational setting, a person does not necessarily need to know exactly what kind of a flower it is we are talking about in order to interpret a sentence that contains a reference to one of them. As long as a person knows the type of thing being talked about, our model of understanding, while not being complete, cannot be said to be wholly impaired or insufficient.

We do not worry about the variability and vagueness of word meanings because we form pragmatic interpretations within the context. Indeed, this does not just apply to isolated words, but to whole utterances as well.

2.2 Meaning in Communication

So we see that there are many ways in which the message that the speaker originally intended to be conveyed to the hearer can fail to be understood and reproduced in the hearer's mind in exactly the same form, and that this need not impede communication of some description from taking place.

2.2.1 Literal vs. Contextual Meaning

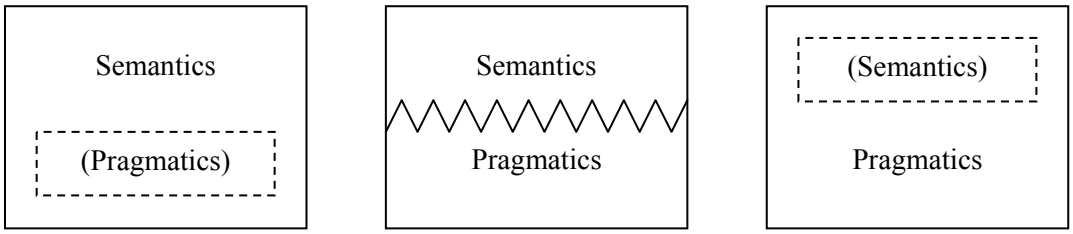
People often make the distinction between speaker meaning and sentence meaning, where sentence meaning is that ground meaning of a sentence taken out of context and which gives the 'real' or 'literal' meaning of the sentence. Proponents of this idea include Lyons (1977: 643) who writes "the meaning of a sentence like 'John is a brave man' is not affected by its being uttered ironically". So if a speaker uses a sentence within a certain context of situation and this sentence can be understood as meaning something significantly different from that of the decontextualised interpretation one might build outside the current situation, then we are

distinguishing between the 'real', conventional, underlying meaning of the words, and the meaning intended by the speaker in the given context.

This idea of the existence of a literal meaning for a sentence (presumably composed of some function of the meaning of the individual words it contains) is a highly controversial and much debated one. For it is unclear that words can ever be taken out of context. Note that discussions on the vagueness of language (briefly touched upon above) can be used as arguments against the literal meaning hypothesis – it may well not be possible to construct prototypical contexts for language at all, or even one and only one interpretation of a sentence. Language is much richer than that. Fodor says: "it is only qua anchored that sentences have content" (1987: 50).

This poses a serious problem when trying to determine the overall meaning of an utterance in speech, a problem that is tantamount to a paradox. For in order to understand what a speaker is trying to say, a hearer must know the meaning of the words; but the meaning itself is subject to contextual re-interpretation. The difficulty with deciding how meaning is determined in language stems from the way it is approached by the two different disciplines of semantics and pragmatics. This dichotomy of meaning has best been defined by Leech (1983: 5-6) as a difference in the use of the verb 'to mean', so that semantics is concerned by the question 'What does *X* mean?', whereas pragmatics deals with 'What did you mean by *X*?'.

Attempts to wholly incorporate meaning into either semantics or pragmatics have all failed at some point or other (I shall be reviewing this topic again in Chapter 4, so I will not go into detail here). Leech (1983: 6) shows how adherents of various theoretical standpoints can be viewed diagrammatically in the following way:



'Semanticism' (Pragmatics subsumed) | 'Complementarism' (Both working together) | 'Pragmaticism' (Semantics subsumed)

Figure 2.1 Divisions in approaches to the ascription of meaning.

Leech argues that the complementary approach, with both the conventional meaning of the words and the contextual interpretation being accommodated, is the most reasonable to assume. While I concur with this conclusion in theory, in practice this still does not solve the dilemma of which 'meaning' should take priority in the interpretation of an utterance.

2.2.2 Variable Interpretation

Language is essentially very flexible, which is what enables the communication of new thoughts from one person to another, and also the application of familiar words to new concepts (an example of this might be the creation of new analogies or metaphors). As such, language is not a fail-safe means of exchanging information but requires considerable effort on the part of speakers to construct intelligible messages, and on the part of hearers to work out what the speaker means. This allows for cases of misinterpretation and error, not only between the speaker and a hearer, but also between different hearers in the interpretation of the same utterance. Because utterances can often be understood in more than one way, there is no real guarantee that your hearer will pick the meaning you intended to convey. This phenomenon is sometimes even exploited by a speaker to communicate different messages to different people by means of a single utterance (an example of this might be when a speaker uses a phrase with a conventionally ascribable meaning as an agreed signal to one of his hearers).

These issues have mainly come to the fore over the past 50 years or so because of a basic shift within the discipline of linguistics from studying formal written texts to looking in detail at spoken language, and in particular the spontaneous, unscripted speech to be found occurring naturally within co-operative conversation. Of course, even written texts are subject to differing interpretations, but this is beyond the scope of this dissertation.

2.3 Adequate Communication

Multiple interpretations of an utterance in context throw up serious problems for those concerned with theories based on the notion of the applicability of truth conditions. The work of Quine (1960) approaches this difficulty by looking at the problems of translating, where he argues that it is impossible to tell what a writer actually intends by the use of a particular word, and so therefore one can only do one's best to interpret a word within the whole text. The same problem can be noted for those who try to paraphrase or summarise other people's concepts (as indeed I have tried to do here), even when the language is shared by both parties – there is no assurance that one has truly understood the original ideas under discussion.

Davidson (1974) suggests that people can only understand language that expresses similar underlying conceptual schemes to those that they themselves share. This seems to me like a rather narrow and restricted view of communication. When taken to its logical conclusion, it fails to account for the fact that we acquire these conceptual structures ourselves as we grow up and are not born with them in place. Accepting such a view would also dismiss the possibility of creativity in our thought, and plainly we are capable of thinking new thoughts and successfully communicating them to others.

2.3.1 Accurate Interpretation of Speaker Meaning

So, the problem remains: a speaker has some meaning in mind when he produces an utterance; so, unless the speaker is completely irrational, the utterance can be said to have a single correct interpretation in the mind of the speaker, even when the meaning of the utterance is ambiguous in the context. Very rarely, as mentioned previously, a speaker will use knowledge shared with one of his hearers in order to communicate different meanings to different members of his audience, but this is not the normal way we communicate. So the question is, how does the hearer know he has identified the speaker's one correct interpretation of the utterance?

Almost no matter what we do, there will always be some doubt as to whether the interpretation of an utterance by the speaker and the corresponding interpretation of the same utterance by the hearer will match and correlate. Fodor says in response to those who insist on there being a 'correct interpretation' to every utterance that it depends on speaker and hearer having identical thoughts:

> …how much (and what kinds of) similarity between thinkers does the intentional identity of their thoughts require? This is, notice, a question one had better be able to answer if there is going to be a scientifically interesting prepositional attitude psychology. ([1986] 1990: 426)

As we noted at the start of this discussion, whatever mechanism it is that is used for communication, it cannot result in neural equivalence between the thoughts of the speaker and the hearer – two minds can never truly be said to think identically. However, it is not necessary to think identically in order to come to the same interpretation of an utterance. Indeed, arguably it may not even really matter. But, how do we judge when the speaker's and the hearer's interpretations are close enough to the 'correct' version that the speaker intended to convey by uttering those particular words?

For example, even in a simple referring phrase 'In some ways, he's very like Margaret Thatcher'[1] it is not always certain that the hearer will have interpreted the phrase correctly, even if the hearer correctly identifies 'Margaret Thatcher' as 'the former British prime minister'. In Fregean terms, he will have, because he has correctly associated the phrase with the individual. But that surely is not enough. The hearer must also have an exact copy of the background knowledge of the speaker concerning Margaret Thatcher in order to be able to interpret what this means correctly. Even if the hearer infers the same attributes as the speaker intends, say Margaret Thatcher's standing up for her beliefs in a particularly aggressive and uncompromising manner, because the hearer believes that this is what the mutual friend under discussion is like, the hearer may still make the wrong assumption about what the speaker

---

[1] This example is taken from Brown (1995: 20).

means. The hearer may think such traits admirable, whereas in actual fact the speaker is criticising their friend for them. Can the hearer really be said to have understood the speaker correctly? If we restrict our view of interpretation to just the correct ascription of sense and reference, then the answer would be yes. But plainly the hearer has failed to understand the speaker's intention in producing the utterance.

It is difficult to know how much of a problem this phenomenon really is in conversation. I believe, in an actual speech situation, this theoretical example would not appear out of context. Either the speaker would explain and clarify his comment, or the previous conversation would make the interpretation obvious, or the hearer himself would be aware of the multiple possible interpretations (the fact that the meaning is ambiguous) and prompt the speaker for an explanation. In any case, I think in these circumstances this kind of misunderstanding would be relatively rare[2].

The idea of a 'correct interpretation' is supported by the work in psycholinguistics, by Johnson-Laird (1990) for example, who argues that if a correct interpretation did not exist for an utterance then communication failures could neither occur nor be rectified. As our models of meaning do 'come apart', there must be a correct interpretation. Again, I do not know how important this is. Misunderstandings of course do occur, but are often caught very quickly in fluid conversation. The repair of a mistaken assumption of meaning is made by the utterance of something like, "Oh, I thought you meant…" or "Oh, so you actually meant… when you said…". Most of the time hearers will arrive at an adequate interpretation of what the speaker meant. On an everyday basis we seem to be able to communicate effectively; we do this mainly by constantly using feedback, paraphrase and summary of what the speaker has just communicated to us ("Are you saying that…?") to reassure ourselves, and by consistency checking ('does what the speaker says now make sense with what he said before?').

2.3.2 Language as Action

Language use is not just an exercise in understanding as it is for example when we are learning to speak an unfamiliar foreign language. It is not used randomly and out of context, but often with a clear aim or as a focus for some further action. It is this context that will resolve for the participants what is an acceptable level of understanding. If a hearer is dissatisfied in some way

---

[2] There are other circumstances where human communication is fraught with potential misunderstanding however. This is particularly true of communication between members of the opposite sex for instance, especially if they do not know each other very well. Ask for separate accounts of how a date went and you might well get significantly different answers!

with his mental model of what the speaker has said, he will ask for clarification, or query his understanding by rephrasing what has been said in his own words.

Not only this, but it is a mistake to think that once an interpretation has been made that it is immutably fixed, and stored in the state it was originally conceived. We do not keep a permanent mental representation of an utterance. We incorporate our interpretation and the various beliefs and knowledge attendant on this interpretation, as well as our understanding of the function of the utterance, into our mental model to inform an appropriate response, and to provide further interpretations of discourse later on. If evidence in later talk reveals that we have misunderstood something earlier on in the conversation, we will refine and update our model accordingly. This updating of the model in the light of new evidence occurs at a localised grammatical level also – garden path sentences are a good example of this. I am suggesting here that similar processing happens on a larger conversational scale.

2.3.3 The Roles of the Speaker and Hearer

A speaker does not form his utterances using the only possible set of words for the 'correct' communication of his ideas, but packages what he says in a way he believes the hearer is most likely to understand in the context of the discourse situation.

If the speaker includes too much detail in his conversation then it becomes boring for the hearer, or the hearer might become overloaded by too much information and so be unable to process it to make a 'correct' interpretation; too little information on the other hand, will lead to ambiguity. Speech is thus constantly balanced between too much and too little information. A speaker is always vying for a hearer's attention and so must try to convey his message as simply as possible. Minimal specification is often the best strategy for speakers to follow. This is often the way that children behave in conversation because they tend to believe that others (especially adults) are already aware of all the background information necessary to decode their message. (In fact this belief in very young children extends to all behaviour – they are incapable of deceit because of the assumption that the other person has complete knowledge of all that they themselves know.) It is interesting to note that minimal specification is often enough, and is easily expanded at need when extra information is required. This is negotiated between the participants in a conversation at the time the need for it occurs. If it becomes apparent that a hearer is unable to understand all that is said, the speaker can easily switch from a strategy of under-specification to that of over-specification (for example when a hearer's background information is inadequate to follow the references being made by the speaker, as in the case of an outsider joining a closely knit group of friends).

Likewise, it is expected that the hearer will try to make sense of what he is hearing and co-operate in the process of communication. Sperber and Wilson (1995: 158) claim that every

utterance comes with a presumption of its own optimal relevance **for the listener**. This is evidently too strong an assertion; it seems obvious that there are some utterances that intrude on the hearer, and whose outcome is of sole benefit to the speaker (anyone who has been approached in the streets by *Big Issue* homeless magazine sellers will understand what this means). From the hearer's point of view, there is no guarantee that, in the end, it will be in the hearer's interests to attend to what the speaker says. Yet we do pay attention to each other when we speak (sometimes!). Brown says:

> It is not necessary to postulate a universal guarantee of relevance to the hearer as the motivation for a hearer paying attention to what a speaker says. (1995: 27)

All one needs to do is to look at the social aspects of communication to find a reason for a hearer's attention. We can explain a hearer's compliance to the demands of consideration of the speaker's utterance by the general elements of co-operative behaviour that govern all human interaction. We work together in life because in the end we as individuals will benefit. The principle of 'if I listen to you now, you may listen to me later' comes into play – this is the fundamental basis of all co-operative activities. One also never knows when what we hear might be of some use.

The view is often expressed in the literature that speakers take the active role in conversation while hearers are merely passive. Clark says: "All that counts in the end is the speaker's meaning and the recovery of the speaker's intentions in uttering the sentence" (1983: 328).

This ignores the case when the hearer was originally the prime mover within a conversation, when it is the hearer who decides what information is most important in interaction (because he himself has first elicited it). This is an important point for my thesis as it motivates the interpretation of speech acts based on prior speech acts. (Note that Clark's quotation is here taken out of context, and that Clark has in later years embraced the dual roles of speaker and hearer in the joint activity of conversation. I shall cover this in Chapter 8.)

Overlooking the importance of both the speaker and the hearer (and only focussing on the speaker's intentions) is as insufficient as trying to assume that the participants share common goals and contexts. Johnson-Laird points out the fallacy of common contexts (or mutual beliefs) for both speaker and hearer – in conversation there are two, one for each participant:

> …the notion of the context overlooks the fact that an utterance generally has at least two contexts: one for the speaker and one for the listener. The differences between them are not merely contingent, but… a crucial datum for communication. (1983: 187)

In fact, this fails fully to cover all the possibilities, for when there is a third observer present to hear the communication between the speaker and the recognised addressee, surely we must then add an extra dimension of context to the conversation.

We can now see the various roles that people can take in a conversation. It will become important to our discussions later to have drawn the distinction between the different 'current' roles or the 'statuses' of participants within a conversation at the time of an utterance, so I will introduce this now.

Figure 2.2 is reproduced from Clark (1996: 14), who distinguishes between participants and non-participants of a conversation. The speaker, addressee and observer (or side participant) are the immediate players in the current action, what Goffman (1976) calls 'ratified participants'. Other listeners are in the category of overhearers, who have no rights or responsibilities in the conversation. A bystander is someone who is openly present but not included, whereas an eavesdropper is someone who is listening covertly without the speaker's knowledge (see Clark and Carlson 1982, Thomas 1986 and Allan 1998, for a more thorough discussion of multiple receiver roles). For the purposes of my research, I will be particularly interested in the ratified participants, rather than in the bystanders and eavesdroppers, although I shall also be attempting to encompass these in a more overarching theory.

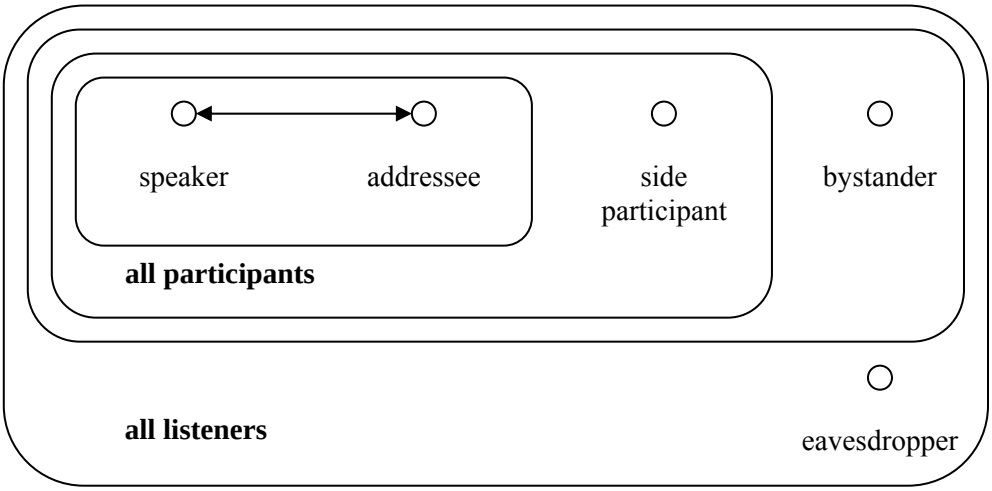


Figure 2.2 Model of participant relationship space.

The idea that different participants will have individual and potentially significantly differing mental models is important to my research. How else can we account for the way that mental models can come apart and misunderstandings take place? The idea of participants sharing a common pool of mutual knowledge and belief is more than slightly suspect. Certainly we make inferences on the basis of what we think we mutually believe, but the most we can ever say is that, as far as we are aware (from the evidence of the conversation), our representation is

consistent and non-conflictive to the best knowledge of our separate belief system. We assume that if it were otherwise, someone would have said something[3].

So, the way that the speaker decides whether their most recent utterance has been correctly understood is to observe the subsequent behaviour of the hearer, since there is no other way of knowing what goes on in a person's mind.

> What listeners have understood from what a previous speaker has said is frequently revealed in what listeners say themselves when they next take a turn at speaking. (Brown 1995: 1)

This is one of the basic assumptions that I will make from the point of view of modelling conversation, and in the assignment of speech act interpretations to transcriptions of spoken interaction.

2.4 Communication in Context

So, having discussed what it is that constitutes communication between human beings, what is the basic contextual information that a participant within a conversation needs in order to hypothesise an adequate interpretation for any given utterance?

(1) *The roles of conversational participants (ratified or otherwise)*: As well as the role of speaker and hearer as I have already discussed in Section 2.3.3, there are a number of roles that participants assume transiently during the course of an utterance.

   (i) **Speaker**: The performer of the current utterance event.

   (ii) **Addressee**: The one towards whom the utterance is directed, and the expected next speaker. Determining who is the intended addressee is not always straightforward, especially when you are deprived of the visual context, when other signals, such as the direction of the speaker's gaze, would often help.

---

[3] Note that this might be an erroneous assumption. Sometimes people do not bother to correct someone's mistaken beliefs, either because they do not care, or in order to avoid being confrontational (with strangers, or with people who have a higher authority).

(iii) **Agent**: In the case of what I shall be calling prescriptive utterances in which some action is being prescribed to some participant, it is important to identify who is the intended agent of the action. In the case of an imperative or interrogative speech act, such as a COMMAND, or a REQUEST, the addressee of the utterance is the intended agent of the prescription; in the case of a self-committing speech act, such as an OFFER, the speaker is the intended agent of the prescription.

(iv) **Hearers (including Side Participants, Bystanders and Eavesdroppers)**: Those other participants who are present in the conversation, but are not expected to contribute. Note that this does not preclude the possibility that one of the hearers will usurp the rightful turn of the addressee. This is a difference that is not often made clear in the literature, but can be of critical significance in my view when it comes to identifying the underlying function of an utterance. Without the theoretical distinction between the intended addressee and other participants of the hearer category, how would it be possible to know when someone is producing an act of COUNTERMAND-ing, for example?

Although I have insisted on distinguishing the roles of the addressee and the casual hearer, and will reflect the unique effects of both positions in my theoretical model, I do not intend to do so in my discussion here. Having made this point, I shall now follow the established tradition of referring to the *speaker* and *hearer* for simplicity's sake.

(2) *The initial assumed common ground or background information*: This is the set of background facts, assumptions and beliefs presupposed to exist in the mental model of the participants. That is to say, all the relevant information that it would be necessary for the speaker and the hearer to have in common in order that the hearer might correctly (or adequately) interpret what the speaker means by a particular utterance. In an ideal world the assumed common ground should always be correct, but often it is not and this is how some misunderstandings occur. The assumed common ground would include such things as the physical and social context, general knowledge, world knowledge and personal knowledge. Often the conversational context is subsumed under this heading, but I have here separated them out into (5) and (6) below. While there is no doubt that there is a very close relationship between these, for my purposes it is essential to consider them separately.

(3) *The actual linguistic representation of the utterance (or verbal act)*: This is the group of words used in the production of the utterance that map onto the total speech act. Following Leech (1983), I include this in this list of elements of context to distinguish between the utterance as an instantiation of a sentence (an abstract grammatical entity) and the use of the utterance within the conversation (whose interpretation then depends on pragmatics).

(4) *The goal (or intended purpose, sometimes called the perlocutionary intent) of the utterance*: Leech uses the terms *goal* or *function* of an utterance in preference to *intended meaning* or *speaker intention* for reasons of side-stepping having to ascribe some sort of positive motivation on the part of the speaker. This is so that we avoid equating a speaker's surface goal with that of his real, underlying intention in producing the utterance (which is entirely unrecoverable from the hearer's point of view). This is a very fine distinction and not really necessary in my opinion. Certainly the speaker intends something to be communicated by an utterance, and the hearer attempts to decipher what the speaker intended. Whether or not there is some hidden layer of intention in the speaker's mind (e.g. to mislead or persuade the hearer in some obscure way) is, I feel, irrelevant. It is the surface 'intention' that we are interested in.

There is a parallel here between this and the identification of illocutionary and perlocutionary acts (which I shall deal with as comprehensively as I can in Chapter 3). Perhaps therefore one should also refer to illocutionary and perlocutionary intentions? So that the act of *assertion* might be the illocutionary intention of an utterance, while the act of *persuasion* is the perlocutionary intention. It is the recognition of the illocutionary intention that is of primary interest to this work, as I would argue that the reconstruction of the speaker's underlying goal is always guesswork for the hearer, informed only by the recognition of the surface intention and by the context.

(5) *The speech act (or co-ordinated social activity) that is being performed*: The function of the utterance in the conversation.

(6) *The current state of activity of the conversational context*: In any rule-governed activity, we maintain a representation of the state of play, which serves as an aide memoir to our current position (so that we can react appropriately in our own turn, and recognise the 'meaning' of the actions of other participants in the activity). In a game of monopoly it is physical representation as depicted by the state of the board, the amount of money in the bank and whose turn it is to go. In the game of speaking, while we do have the physical context to remind us of our current position as regards to physical transactions (Clark 1996), I will be arguing that we also maintain a mental state of play to guide us in our use and understanding of speech acts. Since this is one of the ideas that is central to my thesis I will not dwell too much on it here.

(7) *The list of utterance events (and their speech act interpretation) in the conversation leading up to the current utterance*: Knowledge of the other acts that have been performed in the current conversation (and even that of fairly recent conversations with the same participant) are essential not only for the interpretation of the current utterance, but as a consistency check for (6). If the current utterance does not fit our expectations, we will have to try and

trace the problem back to an earlier state of play to find out if there is some explanation that can account for the hearer being unable to understand this one.

It is the last three of these categories, (5) – (7) that I am primarily concerned with.

2.5 Summary of Issues in Communication

People minimise the risk of miscommunication by judging how much information is needed by the hearer in order to be able to decode the message in context. So a speaker will constantly be deciding whether to maximise or minimise (using pronomialisation or ellipsis for instance) the referents depending upon the status of focus for these referents.

Participants in a conversation will also constantly check whether the message has been correctly conveyed. In spoken language the speaker includes information about how the hearer should treat the content of his utterance, and the hearer will repeatedly feedback reassurances that he has, in actual fact, received the message correctly. Mistakes in understanding are in this way often caught quickly and rectified.

I go along with Brown (1995) who argues for a theory of **adequate understanding** rather than **correct understanding**. She claims that to insist on merely adequate interpretation and on the 'riskiness' of language comprehension is not to say that one cannot ever interpret what someone says 'correctly'. After all, we can and do communicate with each other successfully for the majority of the time. Many utterances are formulated in a conventional or 'formulaic' manner using conventional forms of expression and phrasing in similar frames of context, so our interpretation is quick and apparently seamless. Extended education, learning and training help us acquire this conventional knowledge and apply it in context.

Even so, there is not always a simple solution to an utterance's interpretation; often there are multiple and changing interpretations, the understanding of which is in some cases only partial or incomplete. In these circumstances we make do with the 'best fit' interpretations or ask for more information.

Following on from the contextual desideratum listed in Section 2.4 above, I now turn to look in much greater detail at the origins of speech act theory.

Chapter 3

Philosophical Foundations of Speech Act Theory

> Gastripheres, for example, continued Kysarcius, baptises a child of John Stradling's *in Gomine gatris*, etc. etc., instead of *in Nomine patris*, etc. – is this a baptism? No – say the ablest canonists; in as much as the radix of the word is hereby torn up, and the sense and meaning of them removed and changed quite to another object; for *Gomine* does not signify a name, nor *gatris* a father. – What do they signify? said my uncle Toby. – Nothing at all – quoth Yorick. – Ergo, such a baptism is null, said Kysarcius. (*The Life and Opinions of Tristram Shandy*, Laurence Sterne, Book 4, Chapter 29)

The importance of the theory of speech acts for the understanding of language is widely recognised. "Psychologists, for example, have suggested that the acquisition of the concepts underlying speech acts may be a prerequisite for the acquisition of language in general" (Levinson 1983). If this is truly the case, speech act theory may have a lot to offer the computer scientist seeking a way to 'teach' machines to converse or at least to interact in a meaningful, coherent manner.

It is philosophers of language (such as Wittgenstein and Searle, but especially Austin) who are primarily credited with having noticed that the utterance of some sentences can be treated (under certain well-defined conditions) as the performance of an action. Speech act theory came about as a natural reaction against one of the doctrines prevalent from the 1930s onwards in philosophy – namely that of *logical positivism*. A central tenet of logical positivism was that unless one could in principle verify a sentence (i.e. show it to be true or false), it has to be said to be cognitively meaningless in the strict sense. Taking this view to its natural end, however, one might have to conclude that most written or spoken discourse entirely lacks any cognitive meaning.

3.1 Austin's Theory of Speech Acts

It was the philosopher J. L. Austin who first proposed making a distinction between utterances that could be verified (and were therefore cognitively meaningful according to the definition imposed by logical positivism), and those utterances that may be perceived as performing some kind of linguistic 'act'.

It was amidst concerns about the ambiguity and verifiability of language that were raised by logical positivism, that Austin developed his theory of speech acts. In the series of William

James Lectures, which he delivered at Harvard University in 1955[1] he outlined his objections to the then current theories. He was quite opposed to any theory that would place truth conditions as central to understanding language:

> It was for too long the assumption of philosophers that the business of a 'statement' can only be to 'describe' some state of affairs, or to 'state some fact', which it must do either truly or falsely. Grammarians, indeed, have regularly pointed out that not all sentences are (used in making) statements: there are, traditionally, besides (grammarians') statements, also questions and exclamations, and sentences expressing commands or wishes or concessions.

He also comments that:

> ...both grammarians and philosophers have been aware that it is by no means easy to distinguish even questions, commands, and so on from statements by means of the few and jejune grammatical marks available, such as word order, mood, and the like... (1975: 1)

Austin goes on to argue that basically the problem is founded on a misconception, or misclassification; philosophers were trying to treat all utterances as verifiable statements, when it is clear that many utterances that look like statements on the surface, are either not at all, or only partially, intended to convey any propositional facts to a hearer:

> Along these lines, it has by now been shown ... that many traditional philosophical perplexities have arisen through a mistake – the mistake of taking as straightforward statements of fact utterances which are *either* (in interesting non-grammatical ways) nonsensical *or else* intended as something quite different. (op. cit.: 3)

So there are some utterances, which although traditionally ascribed the grammatical category of 'statement', yet fulfil the following two conditions:

> A. they do not 'describe' or 'report' or constate anything at all, are not 'true or false' [hence problematic for logical positivists]; and
>
> B. the uttering of the sentence is, or is a part of, the doing of an action, which again would not *normally* be described as, or, as 'just', saying something. (op. cit.: 5)

Some examples of these 'statements' are:

---

[1] This series of lectures was posthumously published in 1962 under the title *How to do things with words*, and a later edition was published in 1975. It is this edition that I have referenced throughout this chapter. Therefore please note that all page numbers given hereafter are from this later edition.

‘I declare war on America.’

‘I apologise for hitting you.’

‘I bet you a fiver it will rain tomorrow.’

‘I name this ship the Queen Elizabeth.’

‘I object to your insinuations.’

‘I bequeath you my Van Gogh.’

‘I warn you that trespassers will be prosecuted.’

This type of utterance is used to change the state of the world in some way by performing a kind of action, rather than merely stating something that can be either true or false. Austin called all such utterances (i.e. those containing verbs such as ‘declare’, ‘promise’, ‘object’, ‘pronounce’, ‘name’, etc.) **performatives**; and he called all ‘ordinary’ declaratives (i.e. those that ‘describe’, ‘report’ or *constate* and that can be assessed in terms of truth and falsity, such as the sentence “It is raining”), **constatives**[2].

Austin noticed that performative utterances could ‘go wrong’ as it were, but not by being ‘false’ so much as by being inappropriate, or unhappy. So, for example, in order for the success of a statement such as ‘I now pronounce you man and wife’ in performing an act of (Christian) marriage, it is necessary that the speaker be an ordained minister of the church (i.e. authorised to perform the ceremony/act), that there be at least two witnesses present, that there be a bride and groom, that neither of the afore-mentioned party be already married, etc. If just anyone were to say these particular words in any other context, they would be deemed inappropriate or unhappy, and would fail to bring about an act of marriage. Note, however, that this does not imply that the utterance of this sentence out of context in this way would lack meaning, or be interpreted as gibberish; but just that the act of marriage is not performed. One could very easily envisage a context in which saying ‘I now pronounce you man and wife’ might be taken as an act of joking, or of insulting, for example.

A less formal example of performative utterance **infelicity** would be in an utterance such as ‘I promise to wash the car today’, when the speaker does not do so. Austin says that this cannot be said to be a ‘false’ promise, as the fact that the speaker promised to wash the car would be true. He says that it is the intention to do so that is lacking, so making the act ‘void’ or given in ‘bad faith’. I personally am unsure about this infelicity of intention as described by Austin. My reservations lie in the fact that the speaker’s intentions when uttering this act are actually irrelevant. He might well have intended to wash the car when he made the promise, but for some reason he had to break his word, and the promise remained unimplemented. Thus I would

---

[2] By the end of his book, Austin dismisses this distinction, as I shall explain below.

prefer to describe this state of affairs as a valid act of promising, which becomes invalid only when the content of the commitment made by the speaker to the addressee becomes false (i.e. the commitment that he will wash the car today, where 'he' refers to the speaker, 'the car' refers to some vehicle recognised as the object of the action by both speaker and hearer, and 'today' refers to the date of the utterance – in short, where all items in the utterance have a definite sense and reference, which are understood by both speaker and hearer). The addressee would be justified in complaining to the speaker in such a case regardless of the speaker's original intention. As it will be seen later on, this idea conflicts with, or rather ignores some of Searle's definitions of the preconditions necessary for the success of the speech act of promising.

So, I have explained that according to Austin, there are certain conditions that have to be fulfilled in order for performative utterances to succeed in the context of the conversation: he called these **felicity conditions**. If they are not met, the utterance is said to be **infelicitous**. Austin compiled a list of the different categories of felicity conditions and terminology for their failure, but this is not really relevant to our concerns here. The main point I wished to pick up on was that there are conditions that need to obtain for the success of any speech act (I will say more about this in later chapters).

Having produced this preliminary performative theory, Austin then proceeds to demolish it. He argues that if indeed there is a distinction between constative and performative utterances (i.e. those assessed according to truth conditions and those assessed according to felicity conditions), then it should be possible to isolate what it is that characterises performative utterances. He finds that in the main, performative utterances are typically composed of first person indicative active sentences in the simple present tense. However, this criterion is insufficient by itself, as there are many other examples of this kind of sentence that are not used performatively. E.g.:

'I beat the eggs till stiff.'
'I slam the door thus.'
'I go to the shops (in the morning).'

Often these sentences express a habitual occurrence, or provide an explanation or commentary accompanying a physical action being performed at the same time as the utterance. One way around this problem is to suggest that there is a set of special verbs which are used performatively (although not solely as such), and which can be tested by the insertion of the adverb 'hereby' (or by the presence of the demonstrative/subordinator 'that' immediately after the verb):

'I hereby pronounce you man and wife.'
'I pronounce that you are man and wife.'

So we can see that 'pronounce' here is used as a **performative verb**, whereas, according to the definition above, the following are not:

* 'I hereby beat the eggs till stiff.'
* 'I beat that the eggs are till stiff.'[3]

Austin estimated that there might be well over a thousand such verbs within the English language.

So, with the above mentioned restrictions in mind, surely now we have a means of identifying performative utterances. However, there are further complications; Austin noted that performative utterances could equally well be expressed using apparently non-performative utterances. Consider:

'I find you guilty.' ⇒ 'You did it.' ⇒ 'Guilty.'
'I accept your bet.' ⇒ 'You're on.' ⇒ 'Done.'
'I agree that God exists.' ⇒ 'Agreed.' ⇒ 'Yes.'

On the whole, *given the right situation or context*, all three utterances in each example would comprise the act of convicting, betting or agreeing respectively. Thus, Austin discredits the idea that an utterance can only be a statement of fact (pulling the carpet out from underneath the feet of proponents of logical positivism); he comes to the conclusion that this division between the two types of utterance (constative and performative) is inadequate and misleading. He refines his definition of performative utterances, noting that those that have the correct grammatical format and contain performative verbs are a "*peculiar and special case*" (op. cit.: 63). He calls such utterances with the performative verb named explicitly, **explicit performatives**. The performative verb "makes explicit both that the utterance is performative, and which act it is that is being performed" (op. cit.: 62). So what about non-explicit performatives? Austin makes the following suggestion:

> ...suppose that all performative utterances which are not in fact in this preferred form – beginning 'I *x* that', 'I *x* to', 'I *x*' [where *x* stands for a performative verb] – could be 'reduced' to this form and so rendered what we might call *explicit* performatives. (op. cit.: 68)

---

[3] The symbol * indicates ungrammatical and unacceptable forms of utterance.

Thus, any non-explicit, or **implicit performative** utterances[4] might be rephrased using an explicit performative utterance, with an explicit performative verb. For example:

'It is raining.'	'I state that it is raining.'
'Out!'	'I declare you out.'
'Sixpence.'	'I bet you sixpence.'
'I'll be there at five.'	'I promise to be there at five.'
'Trespassers will be prosecuted.'	'I warn you that trespassers will be prosecuted.'

Therefore we can conclude that the use of explicit performatives may well be merely emphatic: a way of expressing some act as specifically and unambiguously as possible; in contrast to the more common case, where it is the context of an utterance that disambiguates the implicit performative:

> The explicit performative rules out equivocation and keeps the performance fixed, relatively. (op. cit.: 76)

Austin moreover examines what indicators (he calls them "primitive devices in speech") can be found in implicit performatives (i.e. how we can tell which act is being performed when the performative verb is missing):

> The explicit performative formula, moreover, is only the last[5] and 'most successful' of numerous speech-devices which have always been used with greater or less success to perform the same function... (op. cit.: 73)

This seems commonsensical. For in the course of conversation one often finds that the actual act is reported even when it is only implicitly performed. Take, for example, the case when some person *Tom* states some proposition ('God exists', 'It's cold today', etc.), let us call it *X*; then say that one of *Tom*'s hearers, *Dick*, says something like 'yeah' in response. This reply is taken to be an act of agreement with the proposition *X* expressed by *Tom*. If *Dick* contradicts himself later in the conversation, either by saying $\neg X$[6], or by saying *Y*, where *Y* entails $\neg X$[7], then *Tom* might say something like 'But you agreed with me earlier on'. Note that this is perhaps

---

4 Austin used the phrase 'primary utterance' instead of implicit performative utterance, but it is not used by anyone else and so is obsolete.

5 Austin saw explicit performatives as the final unambiguous stage or evolution of language development; this, however, is in my view mere speculation.

6 Following the usual convention, I use the sign "¬" to signify a proposition's negation.

7 For example, if *X* = 'It's cold today' and *Y* = 'It's warm today', then *Y* entails $\neg X$.

just as likely a response in this case than saying 'But you said *X* earlier on'. Such examples indicate that participants in a conversation can and do recognise the linguistic act being performed, even without the presence of an explicit performative verb. I would also note at this point that this recognition is only a postulated or hypothesised one: in actual fact the hearer may have interpreted the act in the wrong way. I will return to this point in later, as it is in my view crucial to the identification of speech acts. However, this is not a subject that Austin touched on.

Austin lists six of these so called 'primitive devices' which conversationalists use in order to identify the implicit performance of an act: mood, tone of voice/cadence/emphasis (prosody), adverbs and adverbial phrases, connecting particles, accompaniments of the utterance, and finally the circumstances (or context) of the utterance. Some of these 'devices' will certainly be relevant to the model I develop later on; so it would be worthwhile looking at these in closer detail.

(1) *Mood*: Austin refers to two uses to indicate overall mood. One is the use of something like the imperative mood (indicated by the main verb) in an utterance, which would typically make it an act of commanding. However, he notes that it may also be an "exhortation or permission or concession or what not!" (op. cit.: 73) – thus, I suppose, showing that it is not a very good indicator of the performance by itself. He gives the example of the phrase 'shut it' in various contexts:

> 'Shut it, do.' *resembles* 'I order you to shut it.'
> 'Shut it – I should.' *resembles* 'I advise you to shut it.'
> 'Shut it, if you like.' *resembles* 'I permit you to shut it.'
> 'Shut it, if you dare.' *resembles* 'I dare you to shut it.'
> 'Very well then, shut it.' *resembles* 'I consent to you shutting it.'

Or again we may use modal auxiliaries:

> 'You may shut it' resembles 'I give permission, consent to your shutting it'
> 'You must shut it' resembles 'I order you, I advise you to shut it'
> 'You ought to shut it' resembles 'I advise you to shut it'

In the model developed later on, I take the mood (grammatical, as defined by Austin) of a sentence to be a marker that legitimises a set (or category) of performative acts. This is not to say that mood is a conclusive or complete indication of the type of act being performed, but that it is a good starting point, especially if the act is performed directly rather than indirectly.

An indirect implicit performative utterance can be described as one that looks as if it is performing one act from its surface structure, but is in fact performing another. For example, the sentence 'The door is closed' seems to be a straightforward statement or assertion of the fact that some door is closed. But if the speaker was carrying a heavy load and required to pass through the doorway, it could be a request, or even an order that the door be opened. Or if it is obvious to the hearer that the speaker is unable to know whether the door is open or not, it could be taken as a question. In both cases, the speaker's tone of voice and the context in which the sentence is uttered are the main indicators of the act being performed. This now leads us to the next 'primitive device' that Austin claimed is commonly used in conversation in place of an explicit performative.

(2) *Tone of voice/cadence/emphasis (prosody)*: These play quite a big part in identifying the role of an utterance in the English language[8]. However, I would have to agree with Austin that there is no adequate way of representing this particular feature in a written transcription of language (which is the medium that I am working in). I would suggest that until we are analysing actual recordings of speech computationally, with some facility or tool to interpret pitch and intonation, this feature for identifying implicit performatives is of no use for studying a transcribed version of speech[9].

(3) *Adverbs and adverbial phrases*: These are used to distinguish the 'degree of strength' (see Section 3.2 on Searle, and Section 10.1.3, for a more detailed discussion) or 'definiteness' of an utterance. For example, 'I'll do *X*' might be made less definite by the addition of the adverb 'probably', or more definite by the addition of the adverbial phrase 'without fail'. So we might characterise by this means the distinction between 'offering' or 'promising' or 'predicting'. The importance of adverbials for identifying different speech acts of the same class, but of subtly differing degrees of strength (or commitment, as I will prefer to call it) is not fully explored by Austin, but is picked up as one of the central themes in Searle's work.

---

[8] Notice that all of these features are specific to English – presumably tone of voice would not play a part in identifying performatives in a tonal language such as Chinese (both Mandarin and Cantonese).

[9] There is a range of pitch tracking systems in existence; the problem is one of interpretation rather than of extraction. There is considerable research into the contribution of intonation to illocutionary force. Hirschberg (1989, 2000), for example, looks at distinguishing questions by their intonation contours. Also, Vassière (1995) looks at the universal features of intonation in language. I do not underestimate the importance of intonation in understanding the force of an utterance, as well as some of its subtler interpretations, but for the purposes of the model that I develop in this dissertation, I will see how far it is possible to get with recourse solely to the words of the utterance themselves. Merging the model of speech act characterisation with models of intonation contours will have to be looked at as a further development.

(4) *Connecting particles*: The use of implicit verbal devices comes "at a more sophisticated level" of performative identification according to Austin. In my opinion, the presence of connecting particles is probably the biggest clue to the force of an utterance in written text. Thus we use 'still' with the force 'I insist that'; 'therefore' with the force 'I conclude that'; 'although' with the force 'I concede that'. However, it's unclear whether these connecting particles are really signalling the force of an utterance, or the performance of an act at a meta-level, to do with conversational moves in the framework of a discourse (see for example Reichman 1985) – so 'although' might indicate any one of the conversational moves of 'indirect challenge', 'sub-argument concession', 'express reservation', or 'enumerate exceptions'. Austin too had reservations about whether 'conclude that' and 'concede that' were performative. I suspect that these 'linguistic indicators/markers' might perform a dual function within the conversation. As I intend to show, speech acts might be seen as the building blocks for conversational moves. However, I do not wish to discuss this point any further at this juncture, as the use of these particles has little immediate relevance to the work carried out herein (although eventually it might prove to be very important in understanding the interplay between different utterances in a conversation).

(5) *Accompaniments of the utterance*: Gestures such as a wink, pointing, a shrug, a frown, a smile, etc. There are many such accompaniments; they sometimes take the place of an utterance completely (e.g. a shrug, can be an eloquent way of expressing lack of knowledge or information for answering a question). Some cultures are extremely rich in expressive gestures of this kind. Unfortunately, even though they are an important way of conveying meaning and force, it is impossible (or inappropriate) to represent gestures in writing without resorting to something like stage directions. As with the case of intonation, I would suggest that until a gesture can be recognised and interpreted in context computationally[10], it should not form part of an analysis of speech acts, even if it is an integral part of a conversation. My justification for this omission is that gestures are not essential for the recognition of implicit performatives. They are not used for example when there are no visual clues available (as when speaking on the telephone) and though some element of meaning or force will be lost, gestures are for the most part replaceable with some substitute phrase (e.g. 'I don't know' instead of a shrug, 'You know' or 'Oh yeah' instead of a wink, etc.); in other words we can get by without them. I am not trivialising the importance of the visual clue in practical communication – it has

---

[10] Gestures and postures are now beginning to be recognised computationally in the field of computational vision (see Watson 1993, Heap and Hogg 1996, to mention two at random) – an important potential application of gesture recognition is the automatic identification of sign language (see Starner and Pentland 1996). A whole new movement in the multi-modal study of communication is now in its infancy (Cassell 2000, Bernsen et al. 2002, Kipp et al. 2002, Serenari et al. 2002). I will return to this briefly in Chapter 10 when I consider the potential future directions of this work.

conclusively been shown to be a significant aid to utterance decoding, e.g. the McGurk effect in McGurk and MacDonald (1976), or other studies such as Massaro and Stork (1998) and House et al. (2001); however, along with a lot of other visual background information (conversational setting, etc.) I do not propose to deal with it as a part of the model I develop for reasons of tractability and scope, but to allow for the inclusion of these elements at a later time.

(6) *The circumstance of the utterance*: This is probably the broadest (and the vaguest) category discussed by Austin. In short, he states that all a participant's background knowledge is used to analyse any utterance. The example he gives is of a report by some person of a past utterance by another person: 'coming from *him*, I took it as an order, not as a request'. In a way, this ability to make the act being performed ambiguous is often exploited by speakers to 'mean more than they say', to insinuate, to intimate, to make an innuendo, and so on.

Austin sees this feature of "vagueness of meaning and uncertainty of sure reception" (op. cit.: 76) as an inadequacy of language. However, I would disagree, suggesting that this is rather an indication of the richness and expressiveness of language, and a problem which must have a solution in context, for we do use this means to communicate. The very fact that there can be this equivocation will sometimes make identifying a unique performative in an utterance very hard indeed. This should not dismay, perturb or put off the attempt to do so. There is no reason why an utterance cannot be classed as more than one performative act should the need arise. A speaker will make a hypothesis about which act is being performed, modifying this hypothesis should it become clear later on in the conversation that the original hypothesis was incorrect. This is often what participants do in a conversation, e.g. 'Oh, I thought you were agreeing with me (but it became clear to me that you were actually performing act *X*)'. It is difficult to say whether this is modifying the original hypothesis, or a complaint that the person who had seemed to agree was being misleading; but I shall come back to this problem in Section 4.5.

Austin sums up his list of 'devices' for identification in the following way:

> No doubt a combination of some or all the devices mentioned above (and very likely there are others) will usually, if not in the end, suffice. Thus when we say 'I shall' we can make it clear that we are forecasting by adding the adverbs 'undoubtedly' or 'probably', that we are expressing an intention by adding the adverb 'certainly' or 'definitely', or that we are promising by adding the adverbial phrase 'without fail' or saying 'I shall do my best to'.

Austin was troubled by the inability to identify consistently performative utterances by any grammatical criterion, but he nevertheless took the view that every performative utterance might be rephrased in the form of an explicit performative, with an explicit performative verb ('state', 'deny', 'order', 'suggest', 'promise', etc.). So he takes a new approach. He reconsiders exactly what he means by "utterance as action" (saying something in order to do something). He defines three separate types of act that are performed by an utterance within a conversation: the

performance of an act *of* saying something (locutionary act), an act *in* saying something (illocutionary act), and an act *by* saying something (perlocutionary act):

(1) **Locutionary act**: "the utterance of a sentence with determinate sense and reference" (Levinson 1983), the act of actually saying something meaningful. By *sense* and *reference* I mean the following. In the sentence 'That is a bank', 'That' is a referring expression, whose referent is that thing (some bank, presumably) to which it refers. 'Bank' has (at least) two different senses: either 'a river bank' or 'a building for depositing money'. Together, the sense and reference of an utterance make up the **content**. So, the sentence could have been uttered in Spanish ('Eso es un banco'), and still the content could have remained the same.

(2) **Illocutionary act**: the making of a statement, offer, promise, etc. in uttering a sentence, by virtue of the conventional force associated with it (or with its explicit performative paraphrase). Each utterance has an underlying **illocutionary force**, which is the function that maps the explicit sequence of words (the locutionary content) onto the illocutionary act. Both the locutionary act and the illocutionary act are therefore crucially dependent on each other:

> To perform a locutionary act is in general, we may say, also and *eo ipso* to perform an illocutionary act (op. cit.: 98)

and you cannot perform an illocutionary act without a locution of some sort.

(3) **Perlocutionary act**: the bringing about of effects on the audience by means of uttering the sentence, such effects being special to the circumstances of the utterance. So, suppose the locution is 'Don't kick the furniture', and the illocution is that of ordering a specific addressee not to kick the furniture (referring by 'furniture' to a set containing items of furniture in the present location). The perlocution could be one of a number of options: it might bring about obedience (the addressee stops kicking the furniture), or defiance (the addressee continues to kick the furniture), or an apology, or anger, or abuse, or a combination of all of these things. I am more than uneasy about calling this an 'act'. Even Austin slips into calling this phenomenon a 'consequential effect'. And yet there is a problem of terminology here because it is an 'effect' that is brought about by the locutionary act (and therefore by definition by the illocutionary act also). My main worry about calling a perlocutionary effect an act is that somehow this conveys the idea that the speaker directly brings it about that the hearer/addressee is convinced, or frightened, or upset, etc. There may be some <u>intention</u> by the speaker to bring about these effects, but he has little or no control over whether this state of affairs takes place. Perlocutions are relatively indeterminate in nature, as can be shown by an utterance of surprise at the unexpected or unintentional effect of one: 'Oh, I'm sorry, I didn't mean to frighten you (... I meant to amuse/inform/chastise/bewilder/excite/etc. you)'. However, there is no doubt that what Austin calls the 'perlocutionary act' has significance for the theory of speech acts, and some account must be made of it in order for a complete theory to exist; for participants in a

conversation can and do recognise these perlocutions: this explains such utterances as 'I've convinced you, haven't I?', 'Don't be alarmed', etc. It remains to be seen whether there is some mechanism by which the performance of an illocution somehow ranges over a discrete set of perlocutions, or whether this set is infinite. To put it another way, it needs to be clarified whether there is a determinate set of perlocutions for every illocution. I have no answer to this last point, and indeed, it is beyond the scope of my research.

It is the illocutionary act that has become synonymous with the **speech act**, and is therefore of particular interest to me. It would seem to me to be the option that holds out the best hope for identification and classification. This is because illocutionary acts are performed by a combination of the content of an utterance and its force (both of which have been described above), so, unlike perlocutionary acts, they are relatively determinate and circumscribed in nature. These are very important implications when considering the formal specification and computational implementation of speech acts.

Austin's solution to the impasse that he reached over the identification of illocutionary acts in their implicit form (as we shall now refer to implicit performative utterances/acts), was to try to categorise, according to their effects, the **illocutionary verbs** found in their explicit form. He displayed a certain amount of trepidation when distinguishing different classes, or families, of speech act, and admitted to being far from happy about the end result, believing that he may well have cross-classified some of them. He has been criticised for being somewhat unsystematic and unprincipled in his approach, but I think that this is unfair considering his own disclaimer[11]. The following is a brief summary of Austin's five classes of illocutionary acts[12]:

> **Expositives**: "...the clarifying of reasons, arguments and communications...". (Austin includes questions in this class.) Examples of these are: 'state', 'deny', 'remark', 'inform', 'ask', 'testify', 'accept', 'correct', 'deduce', 'interpret', 'illustrate', etc.
>
> **Exercitives**: "...an assertion of influence or exercising of power...". Examples of these are: 'order', 'warn', 'bequeath', 'advise', 'nominate', etc.
>
> **Commissives**: "...assuming of an obligation or declaring of an intention...". Examples of these are: 'undertake', 'promise', 'covenant', 'contract', 'swear', 'bet', 'plan', etc.

---

[11] Austin's pupil, Searle (1969) was to modify these classifications and put them on a more principled footing, and he has been accused of being too rigid (I shall describe his work shortly).

[12] Note that I have not dealt with these in the same order as Austin. This is mainly so that Searle's categories and my own all correlate.

Verdictives: "...an exercise of judgment". Examples of these are: 'acquit', 'convict', 'rule', 'estimate', 'value', 'calculate', 'analyse', etc.

Behabitives: "...the adopting of an attitude...". Examples of these are: 'apologise', 'thank', 'sympathise', 'resent', 'command', 'welcome', 'bless', etc. (op. cit.: 163)

The final question to be answered now is, where does all this leave the constative-performative antithesis? Austin gives a subtle twist to his theory by noticing that constative utterances are also subject to certain felicity conditions:

> ...the constative utterance is every bit as liable to unhappinesses as the performative utterance, and indeed pretty much the same unhappinesses. (1971: 19)[13]

The upshot of the argument is that rather than being two separate categories, they are in fact both parts of a more general theory of speech acts. He concludes that utterances in general have the following related features: (1) felicity conditions; (1a) illocutionary force; (2) truth value; and (2a) locutionary meaning (sense and reference). The final analysis he makes of what is needed to advance the development of speech act theory is summed up as follows:

> The total speech act in the total speech situation is the only actual phenomenon which, in the last resort, we are engaged in elucidating. (Austin 1975: 148)

and:

> What we need, it seems to me, is a new doctrine, both complete and general, of what one is doing in saying something, in all the senses of that ambiguous phrase, and of what I call the speech-act, not just in this or that aspect abstracting from all the rest, but taken in its totality. (1971: 22)

We have spent some time on Austin, because many of the issues he raises are still unanswered today, and are clearly of great relevance to my research. His work is the foundation of most subsequent theories of speech acts, not the least of which is that of John Searle, to which I now turn.

3.2 Searle's Theory of Speech Acts

Searle's influential work (1969, 1979) largely followed on where Austin's left off. Much of Searle's earlier work had been an attempt to systematise and formalise Austin's main ideas. In this respect, some claim that he was not altogether successful; it has been argued by Levinson

---

[13] This is from a lecture entitled 'Performatif-constatif' that Austin gave in France in 1958; it is reproduced and translated in Searle (1971).

(1983: 238n) that his analysis was, if anything, less clear than Austin's original. However, in his later work (Searle and Vanderveken 1985) few could fault him on his systematic approach.

Austin's main thrust of argument had been to show that two hitherto unrelated and apparently non-complementary strands of the philosophy of language were necessarily related in a general theory of speech acts (the performative-constative antithesis). Searle also subscribed to this idea (indeed much of Searle's work is very similar to Austin's). His basic assumption is that the speech act is the minimal unit of linguistic communication and he appeals to what he calls the **principle of expressibility** (which states that "whatever can be meant can be said"[14]) to justify his treatment and classification of illocutionary verbs as equivalent to an analysis of illocutionary acts. He claims that the principle:

> ...enables us to equate rules for performing speech acts with rules for uttering certain linguistic elements, since for any possible speech act there is a possible linguistic element the meaning of which (given the context of utterance) is sufficient to determine that its literal utterance is a performance of precisely that speech act. To study the speech acts of promising or apologising we need only study sentences whose literal and correct utterance would constitute making a promise or issuing an apology. (1971: 20-1)

There is a lot of controversy over the soundness of the principle of expressibility as Searle defines it (see Gazdar 1981, Levinson 1983 and Wallis 1995). If it can be shown to be false[15], then Searle's whole motive for studying illocutionary acts by studying illocutionary verbs is defective. I shall return to this point when I look at indirect speech acts in Chapter 4.

Searle differed from Austin in that he was not happy with the distinction Austin made between locutionary, illocutionary and perlocutionary acts. He especially disliked, and indeed rejected altogether, the distinction between the first two of these three (1969: 23n). He saw locutionary acts as constitutive of illocutionary acts, and therefore advocated a rigorous and systematic investigation of the latter alone (which would thereby subsume the former)[16]. Searle had the following observation to make on the point:

> To perform illocutionary acts is to engage in a rule-governed form of behaviour. (1971: 40)

---

[14] Another way of expressing this principle in Gricean terms is that "for any meaning *X* and any speaker *S* whenever *S* means ... *X* then it is possible that there is some expression *E* such that *E* is an exact expression of or formulation of *X*" (Grice 1957, 1968).

[15] The distinction between illocutions and perlocutions, and specifically the indeterminability of perlocutions, seems to belie the principle.

[16] See Holdcroft (1978) for a discussion on this point.

In order to ask a question, make a statement, make a promise, etc., there are a number of "necessary and sufficient" conditions which, when fulfilled, constitute the performing of that illocutionary act. Here Searle points out the crucial difference between *constitutive* and *regulative* rules[17]. Regulative rules are those that regulate activities that are already in use. For example, traffic rules and regulations control the activity of driving, dieting controls the activity of eating, etiquette controls social interaction, and so on. On the other hand, constitutive rules make up part of the activity and cannot be separated from it (e.g. you can still drive a car without following the traffic regulations – people often do – but you cannot drive a car without switching on the engine, pressing in the clutch, putting it in gear, pressing the accelerator, etc.). Examples of a constitutive rule in chess would be of the form "a checkmate is made if the king is attacked in such a way that no move will leave it unattacked" (1971: 41). You cannot play chess without this rule (and others like it) – the rules are the game; thus the rules concerning chess define or constitute chess itself. The importance of ascribing constitutive rules to illocutionary acts is that Searle can make the step of saying that the rules which govern an illocutionary verb (or other illocutionary force indicating devices in an utterance in the case of an implicit speech act) are constitutive of their analogous illocutionary act[18]. Constitutive rules tend to be of the form '*X* counts as *Y*'; thus 'I promise to...' counts as an obligation by the speaker to do the propositional content of the utterance.

Different utterances can often have features in common with each other:

'Mandy will have a drink.'[19]
'Mandy, have a drink!'
Mandy: 'I will have a drink.'
'Will Mandy have a drink?'

These all perform different illocutionary acts: the first would be in general an assertion, the second an order or request (more commonly indicated by the inclusion of preverbal *please*), the third a promise or expression of intention, and the fourth a question. However, along with each different act, there is an element or propositional act that is common to all of the four illocutionary acts given above. In each utterance, the speaker *refers* to a certain person called Mandy, and *predicates* the act of having a drink (in the third case, of course, Mandy performs an act of self-reference). In none of the utterances is this the only act that the speaker performs,

---

17 Searle borrows these definitions from Rawls (1955).

18 For a counter-argument to this conclusion see Gazdar 1981.

19 The original examples given by Searle are variations on 'Sam smokes habitually'.

but in every case it plays a part. Searle calls this common element the (propositional) **content**[20] of the illocutionary act. Thus, the content can be expressed as the phrase *subordinator*[21] + 'Mandy will have a drink', and each of the four utterances above could be rephrased using an explicit illocutionary verb:

'I assert that Mandy will have a drink.'
'I order that Mandy will have a drink.'
'I promise that I [Mandy] will have a drink.'
'I ask whether Mandy will have a drink.'

So, in the utterance of a sentence, a speaker expresses a proposition, and doing so can be part of an illocutionary act. (I say only *can be*, because an expression of pain such as 'ouch' has no propositional content). If the content is the same, then it is the function or force of the utterance that is different. In the same way, we can have utterances with different contents, but the same illocutionary force (or function). For example:

'Mandy will have a drink.'
'Mandy will eat an apple.'
'Maradonna will score a goal.'… etc.

So far, Searle's theory matches up well to Austin's. They come to the same conclusion that the illocutionary act is given by the illocutionary force and the content of an utterance, with the force in effect working as a function on the content. Searle also makes a distinction between the utterance itself and the propositional content; so the utterance of a sentence such as:

'Ms. Amanda Schiffrin will consume a glass of gin and tonic.'[22]

could be said (under certain conditions, and in the right context) to be performing the same *propositional act* and the same *illocutionary act* as the first of the four utterances given earlier, but by means of a different *utterance act*.

From the point of view of semantics, Searle claims that the propositional indicator in a sentence can be distinguished from the illocutionary force indicator (because every sentence can be re-expressed using an explicit formula which – allegedly – contains an illocutionary verb that

---

20 See Gazdar (1981) for his proviso to Searle's definition of content (later taken on board in Searle and Vanderveken (1985)).

21 Subordinators include 'that', 'whether' and 'if', and introduce a subordinate clause.

22 Any similarity between this example and existing persons is purely coincidental.

uniquely identifies the force of the propositional content contained in the rest of the sentence). So, utterances contain two (not necessarily separate) devices: "proposition-indicating" and "function- [illocutionary force-] indicating". Among the list of illocutionary force indicating devices (commonly labelled IFIDs) that Searle gives are: word order, stress, intonation contour, mood of the verb, and the set of illocutionary verbs[23].

Thus illocutionary acts may be represented in the following way:

$$F(p)$$[24]

where F is one of the illocutionary forces (functions) – such as asserting, commanding, promising, or questioning – and p is a variable containing the propositional expression[25]. One of the reasons Searle gives for introducing this formalism is to point out the difference between illocutionary and propositional negation:

Illocutionary negation:	$\neg F(p)$	Not(*Promise*(p))
Propositional negation:	$F(\neg p)$	*Promise*(Not(p))

This distinction is trickier than it seems. For, 'I promise not to p' places the speaker under an obligation to bring it about that he does not do p. But, 'I do not promise to p' removes from the speaker any obligation to bring it about that he does p. Thus 'I do not promise to p' could be seen as a refusal to promise, and not a promise at all:

Illocutionary negation:	$\neg F(p)$	*Refuse*(*Promise*(p))

So illocutionary negation does not ascribe a negative proposition to the speaker, but rather tends to show an ambivalent attitude towards the proposition. Thus 'I do not promise to p' could be rewritten as 'I might do p, but don't hold me to it' or 'I refuse to say/commit myself whether I'll do p or not'[26].

---

[23] See Austin's list given in Section 3.1 for a comparison.

[24] Note that not all illocutionary acts are in this format. E.g. 'Hurrah for Leeds United', which is of the form $F(r)$, where r is a referring expression. For the purposes of this work, I am not interested in these other forms.

[25] In Searle and Vanderveken (1985), the notation is changed to $F(P)$, where P is a proposition that is a function of the meaning of p. Presumably this was in response to criticisms such as Gazdar's (1981).

[26] I have mentioned this distinction here, but have not included it in my model due to a certain feeling of unease. I am not convinced that Searle's (1969) account of illocutionary negation is entirely adequate. Furthermore, I feel that this phenomenon would require deeper investigation before decisions are made concerning how it should be represented.

So, to sum up, Searle discusses the question 'what is an illocutionary act?', and comes up with the following definition: an illocutionary act is composed of an illocutionary force, working upon a propositional content, and that the rules governing illocutionary force are constitutive of their corresponding illocutionary acts.

This has important implications for Austin's original idea of felicity conditions[27]; it leads to the conclusion that they may not just be ways in which illocutionary acts can fail or go wrong, but they might actually be definitive or constitutive of different illocutionary acts. So that, by listing and categorising the felicity conditions of utterances, one would be able to list and categorise different illocutionary acts[28]. Searle gives the example of the felicity conditions necessary for the illocutionary verb *promise* to be 'legal' or 'felicitous' as follows (from Levinson 1983: 239):

(1) The speaker said he would perform a future action.
(2) He intends to do it.
(3) He believes he can do it.
(4) He thinks he would not do it anyway in the normal course of events.
(5) He thinks the addressee wants him to do it (rather than not to do it).
(6) He intends to place himself under an obligation to do it by uttering the sentence 'I promise...'.
(7) Both speaker and addressee comprehend the sentence.
(8) They are both conscious, normal human beings.
(9) They are both in normal circumstances – e.g. not acting in a play [or telling jokes].
(10) The utterance contains some illocutionary force-indicating device (IFID) which is only properly uttered if all the appropriate conditions pertain.

Number (1) is what Searle called the **propositional content condition**, (2) is the **sincerity condition**, (3) – (5) are the **preparatory conditions**, and (6) is the **essential condition** (explained in more detail below). Numbers (7) – (10) are general to all speech acts and so can be ignored for the purposes of this work. Numbers (1) – (6) can be said to be the specific preconditions necessary for, and constitutive of, the act of promising.

So, by analysing the felicity conditions for illocutionary verbs such as 'request', 'assert', 'question' (as in 'ask a question', not as in 'I doubt'), 'thank (for)', 'advise', 'warn', etc., Searle

---

27 Hedenius (1963) argues against analysing performatives in terms of felicity conditions.

28 This is a crucial point, as I will be arguing later that by seeing some felicity conditions as preconditions of illocutionary acts, one might be able to characterise individually any well specified speech act.

simultaneously gives an account of the illocutionary acts of requesting, asserting, questioning (asking), thanking, advising, warning, etc., respectively, defined by the essential condition[29]. Thus felicity conditions are recast as components of illocutionary force; Searle (1969) classifies these conditions into four logical types, which together make up illocutionary force:

Propositional content conditions: These are restrictions that the felicitous performance of an illocutionary act places on the content of an utterance (such as tense or subject). For example, to make a promise, the content must predicate a future action; you cannot say something like 'I promise to have done it by last week' (at least, not while being serious). To apologise, it is necessary that the speaker apologises for his own actions or for some occurrence for which he is responsible (this might include the omission of an action to which he was committed). You cannot apologise for the laws of gravity (once again, at least, not seriously). Violations of this condition are either syntactically or linguistically odd, or instances of non-literal speech acts (such as joking).

Preparatory conditions: These are the presuppositions that the speaker makes about the illocutionary act. For example, if he promises to do something for a certain addressee, it must presuppose that it be in the addressee's best interests for him to do so[30]; if he apologises for some action, it presupposes that this action must be reprehensible in some way.

Sincerity conditions: Illocutionary acts typically express psychological states of mind, thus when I assert *X*, I express my belief in *X*; when I promise to do *X*, I express my intention to do *X*; and when I request, or command, or order that *X* be done, I express my desire or want that *X* be done. Violations of this condition consist of my expressing a psychological state of mind that I do not possess (i.e. lying, or giving a false impression) – for example when I assert without belief, promise without the intention to perform the action, or command without the desire, want, or need for the action to be performed.

Note that the sincerity condition is sometimes openly flouted, for example, in the use of irony. Recognising the speaker's lack of sincerity in this case becomes an integral part of language understanding (I will cover this point in more detail in Section 4.3). This, however, is different to the case when a speaker's internal sincerity is lacking. As I have already argued, I will not

---

29 See Searle 1969: 66-7, for tables of the felicity conditions for these acts and others.

30 An example that seems to disprove this preparatory condition is the use of 'promise' in a sentence such as 'I promise I'll punch you if you say that again'. This is a non-literal use of the illocutionary verb, which would be interpreted as a threat in this particular instance. I will cover this point in detail later in Section 4.2, and in Chapters 9 and 10, where I shall argue that examples of this type indicate that explicit speech acts should be treated in more or less the same way as their implicit counterparts.

take violations of the sincerity condition per se to be relevant in the identification of a speech act's 'felicitous' performance, because a speaker's state of mind is not something that is accessible to a hearer. A hearer can only judge if an utterance is consistent with what he thinks he knows of the speaker, what he thinks he know about the previous conversation and his world knowledge. If what the speaker says does not conflict with the hearer's model of the current conversational situation, then in all likelihood he will accept the act at 'face value', thus expecting the speaker to be sincere. It is only at a later point that the hearer will be able to judge the speaker's sincerity (although, one can also sincerely promise to do something at the time of promising and then forget to do it). In other words, in terms of speech act performance, I do not think it matters if I make a promise with no intention whatsoever of keeping it; if I say "I promise to do *X*" then I am committed to the action in the current context in spite of my intentions.

Essential condition: This specifies what uttering a certain IFID 'counts as'. For example, a 'request' counts as an attempt to get an addressee to perform an action, a 'promise' counts as placing an obligation on the speaker to do an action, and so on.

In Searle and Vanderveken (1985), a further three elements of illocutionary force, itemised below, are added to the four above: **degree of strength of illocutionary point**, **mode of achievement** and **degree of strength of achievement**. Another aspect, that of **illocutionary point**, replaces the essential condition (they are fairly similar in nature).

Illocutionary point: This is the point or purpose of the illocutionary act, which is given internally by the 'type' of the act. So, the point of an assertion is to inform hearers of the way things are, the point of a promise or vow is to commit the speaker to carrying out some action, the point of an order is to attempt to get the addressee to carry out some action, etc. What is meant by the illocutionary point's being 'internal' to the type of illocutionary act is that if the act is successfully performed, then the *point* or purpose of the act is achieved. Thus, if I make a promise to do *X*, I could have many different intentions: to reassure some hearer, to fill in an embarrassing gap in the conversation, to show off, to express irritation, etc. But none of these actually constitutes or captures the essence of making a promise; when I make a promise, I declare a commitment to doing something (in this case, *X*). So, the illocutionary point of an illocutionary act type is the point that is essential to the act's being of that type[31].

Searle and Vanderveken have the following to add:

---

[31] I.e. the illocutionary point indicates or determines the *direction of fit* of the illocutionary act. I shall cover this idea shortly.

> Illocutionary point is only one component of illocutionary force, but it is by far the most important component. That it is not the only component is shown by the fact that different illocutionary forces can have the same illocutionary point as in the pairs assertion/testimony, order/request and promise/vow. In each pair both illocutionary forces have the same point but differ in other respects. The other elements of illocutionary force are further specifications and modifications of the illocutionary point, but the basic component of illocutionary force is illocutionary point. (op. cit.: 14)

Degree of strength of the illocutionary point: Often, different illocutionary acts are performed simply by varying the degree of strength of the illocutionary point. For example, if I *order* someone to do something, this act is stronger than if I *request* that he do it.

Mode of achievement: Some illocutionary acts require special conditions to obtain before their illocutionary point can be achieved. For example, in order to *testify*, a person has to have the authority to do so by being sworn in as a witness; when he makes an assertion, his status as a witness makes it count as a testimony.

Degree of strength of the sincerity conditions: This is analogous to the degree of strength of the illocutionary point. In the same way, different illocutionary acts can be performed by varying the degree of strength of the sincerity conditions. Thus if I make a *request*, I express my wish or desire that the addressee carry out some act; the degree of desire is not as strong as if I were to *beg*, *beseech* or *implore*.

Following on from the idea that illocutionary point gives the purpose of a certain type of illocutionary act, Searle and Vanderveken classify the different types of illocutionary act according to whether the utterance commits the speaker to the way the world is already, or to the way in which he desires that the world be changed. Examples of the former are assertions, agreements, denials, etc. – these are called the words-to-world fit, or the **descriptive**[32] type, because in judging the success of the speech act we determine whether the words fit the way the world actually is. Examples of the latter are such acts as commands, requests, promises, etc. – these are called the world-to-words fit, or the **prescriptive** type, because in performing such speech acts we are expressing a wish that the world be made to fit the words. For example:

'The door is closed.'	(Words-to-world fit)
'Close the door.'[33]	(World-to-words fit)

---

[32] The labels 'descriptive' and 'prescriptive' are introduced as easier mnemonics for these two different directions of fit. Questions (or 'requestives') will be harder to fit into the direction of fit framework.

[33] Notice that the second example could be phrased as a function of the first in the following manner: Addressee, Bring it about that: The door is closed.

Searle and Vanderveken called this feature of language **direction of fit**[34]. They claim that there are "four and only four" possible directions of fit for any utterance. Here is their account of direction of fit (from op. cit.: 53):

> 1. *The word-to-world direction of fit.*
> In achieving success of fit the propositional content of the illocution fits an independently[35] existing state of affairs in the world.
>
> 2. *The world-to-word direction of fit*.
> In achieving success of fit the world is altered to fit the propositional content of the illocution.
>
> 3. *The double direction of fit*.
> In achieving success of fit the world is altered to fit the propositional content by representing the world as being so altered.
>
> 4. *The null or empty direction of fit*.
> There is no question of achieving success of fit between the propositional content and the world, because in general success of fit is presupposed by the utterance.

When classifying illocutionary acts into different categories they list "five and only five" fundamental types, according to direction of fit: **assertives** (equivalent to Searle's (1969) category of **representatives**), **commissives** and **directives**, **expressives,** and **declarations**. These exhaust the alternatives for direction of fit. Searle and Vanderveken summarise their categories, and their arrow shorthand symbols for direction of fit, thus:

> We can summarize the relations between illocutionary forces and directions of fit as follows:
>
> (i) Assertive illocutionary forces have the word-to-world direction of fit↓. The point of an assertive illocution is to represent how the world is.
>
> (ii) Commissive and directive illocutionary forces have the world-to-word direction of fit↑. It is a consequence of the illocutionary point of the commissive/directive illocutions that they create reasons for the speaker/hearer to change the world by acting in such a way as to bring about success in achieving direction of fit.
>
> (iii) Expressive illocutionary forces have the null or empty direction of fit. Nearly all of the words in English which name expressive illocutionary acts (e.g. apologize, congratulate, thank) name illocutions which are expressions of psychological states which have no direction of fit. ...

---

[34] They acknowledge that the expression 'direction of fit' is originally attributable to Austin (1962), although the idea itself is closer to Anscombe (1957).

[35] "'Independently', with the exception of self-referential speech acts where the state of affairs represented is not independent of its representation, e.g. 'This speech act is performed in English'." (op. cit.)

> ...(iv) Declarations have both directions of fit simultaneously, because the point of a declaration is to bring about a change in the world by representing the world as so changed. These are not two independent directions of fit, but one illocutionary point that achieves both. The double direction of fit peculiar to declarations↕ is not to be confused with both directions of fit independently construed (↕ ≠ ↓ & ↑). (op. cit.: 94-5)

Armed with these types and definitions, Searle and Vanderveken go on to develop an extensive taxonomy of speech acts. I shall return to the idea of direction of fit in Chapter 9, as it is crucial and central to the thesis developed in this work.

In this chapter, I have tried to give as full an account of the theory of speech acts as will be required for the discussion to come in the ensuing chapters. I will now turn to one of the main arguments against this approach to language, namely the phenomenon of indirect speech acts, and outline the various means of solving the problems thrown up (by including the conversational context in the analysis).

Chapter 4

Issues in Speech Act Theory

Out of the frying pan, and into the fire. (*English proverb*)

The position of Austin and Searle as described in the previous chapter, can be contrasted with the attempt to reduce speech act theory to a semantic phenomenon (what Leech calls the semanticist approach – see Section 2.2.1 for details). This is characterised by the view that every implicit performative utterance has some underlying explicit performative structure, and can be rephrased as such without affecting its meaning. This hypothesis is known as the performative analysis of speech acts.

4.1 The Performative Analysis

There are mixed views on this analysis of performative utterances; I wish to give a brief account here of the main arguments against this account of speech act theory. I will not go into great detail, as this has already been more than adequately covered by Levinson (1983: 243-63). Since I reject these arguments for a truth-conditional semantic approach (along with Levinson 1983, Gazdar 1981, and others), I only wish to mention a few of the arguments and their refutations.

Hedenius (1963) suggests that explicit performatives might simply be said to be true. Thus a sentence such as 'I warn you that the bull is going to charge' would be true at the moment of utterance – it is a warning that the bull is going to charge. This seems a radical attempt to circumvent the need for a pragmatic theory at all. Under this theory, implicit performatives are either elliptical explicit performatives, or have as their highest clause in deep underlying syntactic structure a representation containing an explicit performative prefix, whether it is 'visible' or not. This would seem to make sense of some aspects of anaphoric reference, which are otherwise unclear. So the use of the reflexive pronoun 'myself' can be explained in just the case when the sentence:

'The kitchen was cleaned by Bob and myself'

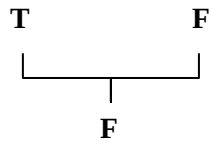
has an underlying sense of the following:

'I tell you that the kitchen was cleaned by Bob and myself'

This way of accommodating performatives is called the **performative analysis** (Ross 1970, Sadock 1974 and Lakoff 1972, 1975 are among its chief subscribers).

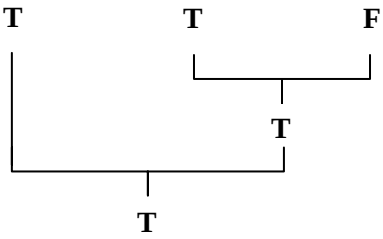
The problem with this analysis is that utterances such as 'Snow is green' would be seen as true in the instant when it was uttered (because its underlying structure is of the form 'I state that snow is green'). Clearly, this is a ridiculous standpoint; howsoever we may analyse the latter sentence, the former, given the way the physical world is, is unarguably false. Lakoff (1975) suggests that the solution to this problem might be to say that both the performative and the compliment clauses have to be true in order for the whole utterance to be true. Thus:

'I state that... snow is green'



This seems counter-intuitive, but does give the same truth values for both sentences. However, a sentence such as 'I stated that snow is green' does not contain a performative use of the verb 'state' and it can be seen that the utterance would be true in the case that the speaker had sometime in the past stated that snow is green. So the analysis of this sentence would be:

'I state that... I stated that... snow is green'



But here it can plainly be seen that the non-performative use has <u>different</u> truth conditions from the performative use. So, this approach has actually *failed* to reduce performatives to truth conditional semantics[1], and we are forced to the conclusion that speech acts are firmly in the realm of pragmatics[2]. As Levinson (1983: 246) puts it:

> …illocutionary force constitutes an aspect of meaning that cannot be captured in a truth-conditional semantics.

However, the argument against the outlook of generative semanticists on speech act theory is not the greatest problem to be faced. One of the biggest difficulties for proponents of speech act

---

1 This is not the complete argument, but I hope captures the main gist of it.

2 Though I shall be adopting Leech's (1983) complementarist view, which comprises both semantics and pragmatics, claiming that in some way both have a hand in the ascription of utterance meaning.

theory (from both the semantic and pragmatic point of view) is the phenomenon known as **indirect speech acts**.

4.2 The Problem of Indirect Speech Acts

The concept of indirect speech acts can only make sense on the assumption of what Gazdar (1981) calls the literal meaning hypothesis, and Levinson (1983) calls the literal force hypothesis. This holds that there is a conventional tie between the form of a sentence and its force. An indirect speech act is one that has the surface syntactic structure of one sentence type, but performs within a conversation as if it were of another. Thus the sentence

'He's going to the theatre'

can be conventionally analysed as a *declarative* sentence type, and would be assigned an assertive speech act such as 'assert', 'state', 'inform', or 'say'. However, under certain circumstances, exactly the same sentence would actually be performing an interrogative role in the conversation, and would have to be interpreted as a question equivalent to 'Is he going to the theatre?'. The same sentence might even be playing a dual role; for example, it may be informing a hearer of some person's (referred to by 'he') plans for the evening, and it may also be an insistence to that person that this <u>is</u> what he will do whether he wants to or not. This form of double purpose is probably most often seen when someone is in authority – especially parents with children:

A: And what are you going to do tonight, Johnny ?
B: He's going to have a bath [aren't you?]

These are not special cases, however, and there is an indefinite number of indirect speech acts. It is not a rare occurrence that participants in a conversation will mistake, or even be unsure of the intended speech act, and request confirmation regarding the correct interpretation:

A: You will go tomorrow.
B: Is that a question, or an order?

This possible ambiguity is often exploited for various reasons; for example in order to make a joke:

A: I'll cook tonight.
B: Is that a promise or a threat?

Or to avoid refusing a request:

A: Is the rubbish out?

B: No.

(In this case, if the indirect approach fails, generally the request will be rephrased, although it might still be in an indirect form: 'Could you put it out please?'.)

All this goes to show that contrary to Heringer (1972: 6) who says that "a given utterance under normal circumstances must be the performance of one and only one illocutionary act", an utterance does not have one fixed illocutionary act interpretation. Searle (1969) recognised that this is the case.

> ...it is important to realise that one and the same utterance may constitute the performance of several different illocutionary acts. There may be several different non-synonymous illocutionary verbs that correctly characterise the utterance. (op. cit.: 70)

Searle not only understood that an utterance could (generally) be interpreted as more than one illocutionary act, but also that it could mean different things to different hearers. He attempts to incorporate indirect speech acts into his taxonomy by claiming that in performing an indirect speech act, one is also performing a literal or direct speech act. So, for the utterance

'Can you pass the salt?'

Searle asserts that there are two acts: a primary act of requesting, and a secondary act of asking, with the first 'built on' or 'reliant on' the second. For Searle all sentences that have interrogative syntactic structure, are said also to have the illocutionary force of a question.

> In cases where these sentences are uttered as requests, they still have their literal meaning and are uttered with and as having that literal meaning. I have seen it claimed that they have different meanings "in context" when they are uttered as requests, but I believe that is obviously false. (Searle 1975: 69-70)

On this point, I would have to disagree with Searle. There seems to be empirical evidence that people do not recognise such utterances as questions at all; or if they do, only as a secondary additional feature (see Gazdar 1981 for details). This explains the way that answering an utterance such as 'Can you pass the salt?' with 'Yes', without the accompanying action of passing the salt as well, is considered to be a (very feeble) joke. Here, the 'indirect act' is the question and not the request, hence the pun.

The problem with this admission is clearly expressed by Schegloff (1976: E3).

> Even when an utterance is in the linguistic form of a question, and seems to be doing questioning, the latter will not be adequately accounted for by the former. For if the question form can be used for actions other than questioning, and questioning can be accomplished by linguistic forms other than questions, then a relevant problem can be posed not only about how a question does something other than questioning; but how it does questioning; not only about how questioning is done by non-question forms, but how it gets accomplished by a question form.

What is it that determines a question? We can see that Searle's insistence that (the literal force of) every utterance with interrogative syntax is an act of questioning, is fallacious. Even so, the phenomena I have discussed so far are in my opinion not good enough evidence to discard the literal force hypothesis completely. If we look at the sentence 'He's going to the theatre' again, we can see that this has a surface syntactic form *declarative*. But illocutionary force is not simply determined by sentence type. This explains why it can function as more than one kind of speech act:

> QUESTION: He's going to the theatre?

Here intonation and stress play a big role. Intonation could be seen as a type of function on the surface syntactic form of a sentence. Thus we have: Y/N-question-indicating-intonation (declarative-syntactic-structure (content))[3].

In the case when this utterance performs more than one speech act, the interpretation depends upon the **conversational context**. In order for the sentence to be interpreted as insistence by the subject of the conversation, it is necessary that the speaker have already requested/ordered that he should go to the theatre, and that the subject has declined. This is a very important point for my research, and provides a basis for speech act identification within the conversational context. I shall return to this analysis of speech acts in context in the following chapters.

Gazdar (1981) and Levinson (1983) use a sentence such as 'May I remind you that your account is overdue?' to show that Searle's position (and, as a consequence, the idea of a conventional link between force and sentence type) is untenable. Their interpretation of such an utterance under Searle's theory would be a request for permission to remind. Levinson argues that this cannot possibly be a request for permission to remind, because reminding is carried out by the expression of the utterance without the permission to do so ever having been granted to the speaker.

I subscribe to an idea of Labov and Fanshel (1977) that the addition of a modal verb at the beginning of such an utterance is a type of *mitigator*, and is tied up with the concept of

---

[3] This would seem to indicate that there might be some form of hierarchy of IFIDs, so that the presence of one particular type over-rides the presence of another.

politeness. Thus, the sentence above is actually simply an explicit speech act of reminding, and can be rephrased as one by excluding the mitigator.

'I remind you that your account is overdue.'

This use is only relevant when the second verb is one of the 'performative' or 'illocutionary' verbs. It is mainly a method to avoid the impoliteness of a direct reference to something that is either socially embarrassing or unpleasant, in order to save the 'face' of the addressee.

The idea of politeness mitigators can be extended to explain a wide number and variety of indirect speech acts; especially those in the syntactic form of questions, that are indirectly interpreted as requests, request-permission and request-confirmation. Often the request is present at some level in the sentence structure in the imperative form. For example, embedded in the sentence 'Can you pass the salt?' is the imperative form 'Pass the salt'. The presence of the modal 'Can' serves to mitigate the force of a direct command, as it is impolite to order people about.

By accepting this account of indirect requests, we solve many of the trying problems of these types of indirect speech acts:

'I don't suppose that you would by any chance be able to lend me some cash, would you?' ⇒ 'Lend me some cash'

'Can you pass me the ice-cream please?' ⇒ 'Pass me the ice-cream'

'You couldn't close the door, could you?' ⇒ 'Close the door'

Admittedly, there are some elements and shades of meaning that are lost by taking this approach; but in general, in the interpretation of these types of sentences people do strip off these politeness mitigators to get at what the speaker is really talking about or asking for, in order to respond accordingly.

However this theory, while accounting for a good many indirect utterances, cannot explain them all. For example, what can we make of an utterance such as:

'Would you mind turning the music down please?'

Clearly, this is a request to turn the music down, but there seems to be no syntactic surface feature in this sentence to indicate an imperative reading. Intuitively, one would say that the reason why this request is framed in such a way is for reasons of politeness, but why use the verbal inflexion *-ing*? One possible answer is that 'Would/Do you mind' is used as an idiomatic phrase indicating a request.

I am aware that I have only skated over the surface of the indirect speech act phenomenon. In common with many (Grice 1975, 1978, Leech 1980, 1983, Brown and Levinson 1978, 1987, Smith 1991), I feel that the driving forces behind a large proportion of our use of indirect utterances are some over-riding principles of politeness (motivated by issues of power, authority and confrontation), which govern language in a social context.

I would agree with Gazdar (1981) and Levinson (1983), who draw the conclusion that there are no literal forces governing the identification of speech acts: no sentence has a conventional literal force that can be extracted from its surface syntactic structure (as defined by Searle 1979), and therefore, there are no indirect speech acts. This means that illocutionary force is almost entirely pragmatic in nature, with neither sentence-form nor sentence-meaning having a fixed bearing on it (i.e. there is no one-to-one correlation between illocutionary force and the form or meaning of a sentence). This does not mean that there is not an underlying propositional meaning, but that this meaning is only fully interpretable in context. Having concluded this, I would say that the sentence form of an utterance must play some role in the identification of illocutionary force; even if perhaps only by restricting the range of speech acts that the sentence can perform, rather than uniquely identifying the literal force of an utterance.

The question remains: how do we explain those types of utterances (hitherto called indirect speech acts) that do not seem to fit the framework for identifying speech acts which I have described so far? The answer seems to be that there is a more general problem: that of mapping illocutionary force onto sentences in the conversational context. In order to be able to search for such a mapping, there are a number of basic principles and assumptions that participants in a conversation must hold.

4.3 Grice's Principles and Maxims for Coherent Conversation

Searle (1969, 1979, 1985) based his analysis of speech acts on the felicity conditions that had to obtain for the success of any particular speech act. This approach holds up fairly well; yet often when the felicity conditions of one speech act are not completely fulfilled, rather than signalling the failure of the speech act, it indicates that some other (what we have called indirect) speech act is being performed. A lot of English jokes, metaphors, and uses of irony are founded on the failure of some of the felicity conditions (e.g. the sincerity condition in the case of irony).

The co-operative principle (and related maxims) of conversation on which such an interpretation of speech acts may be based is mainly attributed to Grice (1975, 1978). This forms the basis of Grice's theory of *conversational implicature*. Grice presents the co-operative principle in the following terms:

> Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. (op. cit.: 45)

There are a number of conventions, or maxims that are associated with this principle, which are (taken from Brown and Yule 1983: 32):

Quantity: Make your contribution as informative as is required (for the current purposes of the talk-exchange). Do not make your contribution more informative than is required.

Quality: Do not say what you believe to be false. Do not say that for which you lack adequate evidence.

Relation: Be relevant.

Manner: Be perspicuous. Avoid obscurity of expression. Avoid ambiguity. Be brief (avoid unnecessary prolixity). Be orderly.

An essential additional maxim to add to those above would be *Be polite*. Grice does not claim that this list is exhaustive, but that it provides the foundation for coherent communication. We all make these ground assumptions about any utterance, and tend to follow the same ground assumptions when producing an utterance ourselves. The relationship between these principles and maxims and Searle's felicity conditions can be clearly seen. However unlike Searle, Grice, in describing the normal underlying 'rules' that operate in conversation, investigates how a speaker can convey different meanings of an utterance by **flouting** one or more of these 'rules'. Generally the flouting of one of the co-operative maxims produces an additional meaning as well as the literal meaning of an utterance and this is the conversational implicature. Grice (1975: 51) gives the following example:

A: I am out of petrol.
B: There is a garage around the corner.

In this conversation, it appears that **B**'s utterance is irrelevant, and therefore in violation of the maxim of relation. But by the co-operative principle we assume that the speaker is trying to convey some relevant information; thus the implicature derived would be that the garage around the corner would be selling petrol. So this is not merely a statement of fact, but a helpful piece of information to **A**, relevant to **A**'s previous comment (notice that the recognition of this implicature is dependant on some background knowledge of the world).

Brown and Yule (1983: 33) have the following comment to make:

> …implicatures are pragmatic aspects of meaning and have certain identifiable characteristics. They are partially derived from the conventional or literal meaning of an utterance, produced in a specific context which is shared by the speaker and the hearer, and depend on the recognition by the speaker and the hearer of the Co-operative Principle and its maxims... [they] must be treated as inherently indeterminate since they derive from a supposition that the speaker has the intention of conveying meaning and obeying the Co-operative Principle.

Grice's theory of implicatures is not directly relevant to my thesis, as I am not dealing with the content of utterances or the inferences performed by the hearers to derive the meaning of an utterance. What I wish to show by describing some of the details of Grice's theory is that the fundamental assumption made by participants in a conversation is that the utterance of a sentence by a speaker will be coherent within the context of the conversation. We assume that speakers are co-operating with us on some level in order to communicate, until we have exhausted all possible interpretations of a sentence. Only then will we draw the conclusion that the speaker is being incoherent and not making sense.

4.4 Leech's Principles of Pragmatics

Leech (1983) repositions speech act theory (Searle's approach in particular) in terms of Gricean conversational implicatures, and argues for the need to separate semantic issues from those that belong firmly in the realm of pragmatics. He does this by developing a set of eight postulates that defines the differences that characterise the 'formal' (semantic) and 'functional' (pragmatic) paradigms, and that emphasises the rhetorical[4] nature of speaking.

4.4.1 Semantics vs. Pragmatics

P1: THE SEMANTIC REPRESENTATION (OR LOGICAL FORM) OF A SENTENCE IS DISTINCT FROM ITS PRAGMATIC INTERPRETATION.

This is the principle that semantics and pragmatics are separate areas of study and that neither can be assimilated into the other. This is what Leech calls the complementarist point of view (as opposed to the semanticist or pragmaticist perspective). I have already covered this in Section 2.2.1.

---

[4] Leech uses the word 'rhetorical' to mean goal-oriented within a speech context, when the speaker uses an utterance with the idea of producing a certain effect on his hearer.

P2: SEMANTICS IS RULE-GOVERNED (= GRAMMATICAL); GENERAL PRAGMATICS IS PRINCIPLE-CONTROLLED (= RHETORICAL).

This is the distinction between the rules of semantics being constitutive of language as contrasted with pragmatic principles being regulative of effective communication. It might be arguable that grammar itself is just a useful framework on which to hang words, and that even an utterance that does not seem to be grammatical on the surface will nonetheless be intelligible within the current context. Note that this is different to the claims of Searle (1969: 38) who says that, "language is a matter of performing speech acts according to systems of constitutive rules". As I noted before, Searle is keen to be able to ascribe one and only one speech act per utterance. There is very clear evidence in speech to contradict this.

P3: THE RULES OF GRAMMAR ARE FUNDAMENTALLY CONVENTIONAL; THE PRINCIPLES OF GENERAL PRAGMATICS ARE FUNDAMENTALLY NON-CONVENTIONAL, I.E. MOTIVATED IN TERMS OF CONVERSATIONAL GOALS.

This postulate is harder to uphold than the two previous ones, P1 and P2. Pragmatic constraints and features become fossilised in the grammar over time, and in so doing, become conventional in use. In other words, pragmatic uses of words or phrases have a tendency to cross over to become grammatical entities (part of the lexicon), so that the boundary between grammar and pragmatics becomes blurred and is often difficult to draw with any certainty. The strongest expression of this postulate therefore can be made along the lines of "grammar is primarily conventional and secondarily motivated; pragmatics is primarily motivated and secondarily conventional" (op.cit.: 29-30). I will be revisiting this idea at the end of this chapter, because it has important consequences when it comes to discussing the process of inference involved in interpreting utterances.

P4: GENERAL PRAGMATICS RELATES THE SENSE (OR GRAMMATICAL MEANING) OF AN UTTERANCE TO ITS PRAGMATIC (OR ILLOCUTIONARY) FORCE. THIS RELATIONSHIP MAY BE RELATIVELY DIRECT OR INDIRECT.

Leech redefines indirectness as a natural consequence of conversational implicatures, which attribute mental states and attitudes to the speaker. He claims that there is no need for extra illocutionary rules to account for indirect illocutions, and in fact there may be no need to make the distinction between directness and indirectness at all. "All illocutions are 'indirect' in that their force is derived by implicature" (op. cit.: 33). This leads on to the idea of a 'default'[5] interpretation of an utterance, which is the meaning that is applied to the utterance in the

---

[5] Bach and Harnish also use the idea of default interpretation in their theory – I shall be covering this in more detail later in this chapter.

absence of any evidence to suggest otherwise. If the default interpretation is inadequate in context, then another meaning is suggested by appealing to one or more of the maxims for co-operative communication.

One criticism that can be levelled at the application of this idea is that neither Grice nor Leech indicate how the maxims are to be used in order to do this, whether for the interpretation or generation of utterances. Which of the maxims for co-operative and polite conversation should you give precedence to? In what order are they to be considered or applied? What contextual constraints are applicable? There is no rigorous model presented to account for the way the maxims are actually used in communication. In essence, there is an excellent explanation of why we use co-operative principles in spoken language, but not of how.

One of the reasons that it is hard to come up with a definitive order of importance for Grice's maxims of co-operative conversation is that the weight one gives each maxim is at least partially determined, if not by an individual's personality, then by their culture. For example, the Argentinians in general feel that the loss of face shown by admitting that they do not know something is worse than the consequences of making something up that is plausible, and for which they may have anecdotal evidence, but for which they have no real proof of veracity. They are not a nation of liars, but have a preference for the maxim of informativeness (quantity) over the maxim of truthfulness (quality). Their mental attitude is one of, 'something is better than nothing', even if that something turns out to be false.

Leech also claims within the scope of this postulate that Searle's speech act theory can be completely reinterpreted in terms of Gricean conversational implicature (for example, the sincerity condition might be said to correspond to Grice's maxim of quality).

In conclusion, he makes the following assumption about pragmatic force:

> if *s* means *F* by *U*, then *s* intends *h* to recognize the force *F by way of the sense (ie: the grammatical meaning) of U*. (op.cit.: 35)[6]

Central to this postulate is that the meaning of an utterance is a function of the sense of the words in an utterance and the utterance's force. Leech concludes that if the meaning cannot be so determined, then it falls beyond the scope of pragmatics. He gives the example of a card player using some sentence such as 'My aunt has a villa in Vladivostok!' as a means of communicating a meaning in code to his partner about the contents of his hand of cards.

On the face of it, this seems like a sound distinction to make, although I could see a way of analysing this as 'X standing for Y', which would then be interpreted in one way by the hearer

---

6 Where *s* = speaker, *h* = hearer, *F* = force and *U* = utterance.

in the know, and as a normal assertion by the other players. So, in a way, even an utterance such as this one has pragmatic force in context. This is a bit of a tenuous example, as I submit that if a card player did say something so incongruous in the context of a card game he would almost certainly be suspected of cheating! Oddly enough, probably because of pragmatic expectations about the kind of conversation one would expect to occur in a game of cards.

Leech claims that there are other situations that are beyond the scope of pragmatic concern. These are where miscommunication occurs, such as when two people do not share the same language, or when some noise interferes with an utterance, or a listener mishears what is said, or when the speaker is being unco-operative, or not observing goal-oriented behaviour for some reason, and so on. I would disagree that this is no business of pragmatics; surely it is mainly because some pragmatic principle is not being observed that the speaker will recognise that the communication is not working in some way?

P5: GRAMMATICAL CORRESPONDENCES ARE DEFINED BY MAPPINGS; PRAGMATIC CORRESPONDENCES ARE DEFINED BY PROBLEMS AND THEIR SOLUTIONS.

Pragmatics is all about problem-solving. A speaker's problem is how to convey a message by the use of an utterance; a hearer's is finding out what is the speaker's most likely motivation for having produced a particular utterance (bounded by considerations of politeness, power, situation, etc.).

Leech represents this problem-solving nature of language in terms of goals and states. His diagrammatic representation of the utterance of the phrase, 'Cold in here, isn't it?' is shown in Figure 4.1 (from op.cit.: 38):

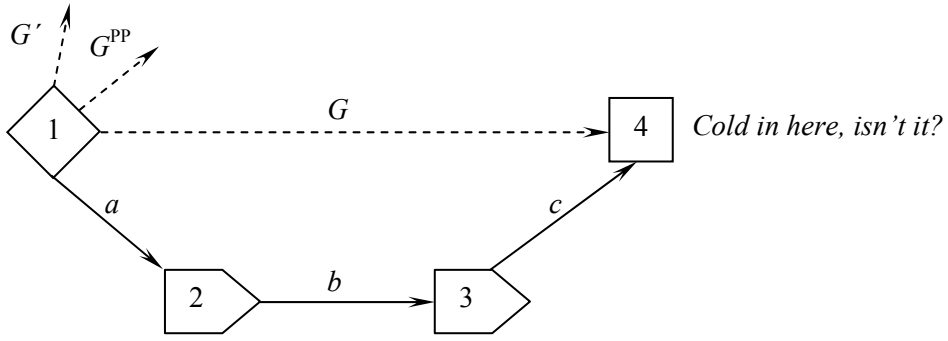


Figure 4.1 State transition diagram of the utterance 'Cold in here, isn't it?'.

KEY to Figure 4.1

	1	=	initial state (speaker feels cold)
	2	=	intermediate state (hearer understands that the speaker is aware that it is cold)
?	3	=	intermediate state (hearer understands that the speaker wants the heater on)
	4	=	final state (speaker feels warm)
	G	=	goal of attaining state 4 (getting warm)
	G^{PP}	=	goal of observing the Politeness Principle
	G′	=	further goal(s) (unspecified)
	a	=	speaker's action of remarking that it is cold
?	[*b*	=	speaker's action of telling the hearer to switch on the heater]
	c	=	hearer's action in switching on the heater

Action *b* is here included as a concession to Searle, who analyses indirect speech acts as secondary acts that contain the underlying meaning of the utterance; Leech argues that this is an unnecessary stage, and that it is actually the hearer's action of inference that occurs. So we can replace the definition above with:

b	=	hearer's action in inferring that the speaker wants the hearer to switch the heater on

I would claim that even this is too strong an analysis of the speaker's utterance (which is why I have marked the state and action with whose interpretation I disagree with a question mark '**?**' above). In fact the hearer cannot infer that the speaker wants the heater to be switched on from the production of such an utterance; that's far too specific a deduction. I would say that the most a hearer can infer is that the speaker wishes to be made warmer, because feeling cold is not a pleasant state in which to be. So, because the hearer is a co-operative and socially well-adjusted person, he forms a plan to bring it about that the speaker is warmer by means of switching on the heater (or lending the speaker a jumper, or closing the window, etc.). The point is that this utterance would rarely if ever be produced out of context, which would clarify what the speaker actually meant to convey in his utterance.

The hearer himself would possibly also wish to make sure that he has correctly interpreted the speaker by checking in some way, perhaps by asking, 'Shall I turn on the heater?'. It would actually be seen as quite odd behaviour if the hearer simply turned on the heater in response to the original utterance without some kind of clarification and reassurance sequence. In a real conversation, getting at the meaning of an indirect utterance is not as difficult as all that; participants can always just ask.

Leech describes the hearer's task as that of heuristically testing all the possible hypotheses for the interpretations of an utterance against the contextual evidence in a repeated cycle. This is shown diagrammatically in Figure 4.2 (from op.cit.: 41).

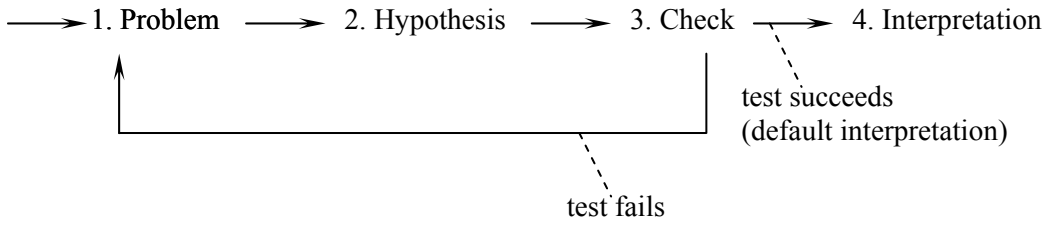


Figure 4.2 Chain of hypothesis testing in utterance understanding.

Leech recognises that this is an oversimplified representation, and fleshes it out by looking at the case of a simple assertion (it does not matter what it is) in the following way (where *s* = speaker, *h* = hearer, *P* = sense(*U*) or proposition, and *U* = utterance):

A.	*s* says to *h* [that *P*]	**Grice's Maxim?**
B.	*s* means [*h* to be aware [that *P*]]	
C.	*s* believes [that *P*]	QUALITY
D.	*s* believes [that *h* is not aware [that *P*]]	QUANTITY
E.	*s* believes [that it is desirable [that *h* to be aware [that *P*]]]	RELATION

A = DEFAULT INTERPRETATION
B = MINIMUM ILLOCUTIONARY ASSUMPTION
C-E = CORROBORATIVE CONDITIONS

If any of the corroborative conditions are not met, the default hypothesis of the assertive force of the utterance will fail, and some other pragmatic hypothesis for the possible interpretation of the utterance will have to be sought.

The way it is here described, the process of finding out the meaning of an utterance might seem quite labour intensive, however Leech suggests that the context may play a role in limiting the potential interpretations:

> …one should not expect that the default interpretations are the same in different contexts. The expectations of addressees will vary according to situation, so that what may be a default interpretation in one context will not be so in another. (op.cit.: 43)

This is certainly true, and it is part of the object of this dissertation to show how it might be possible to constrain the choice of speech act from the conversational context. I also believe that the conventional use of cue phrases and other grammatical constructions may have a lot to do with the speed in which we are capable of finding an adequate interpretation.

It is difficult to understand at this point what Leech is advocating; it cannot be possible that for every assertion made, all the information given above (*A*-*E*) passes through a hearer's mind before he makes an answer (unless this happens at considerable speed at an unconscious level). If it did, I do not believe that we would ever manage to communicate as quickly and efficiently as we do. Perhaps a better way of thinking about a hearer's position towards a speaker's utterance might be to take an action approach such as that expounded by Clark (1996)[7]. This analyses the interaction from the hearer's point of view, and claims that the hearer is more likely to be thinking about what he should do next as a response to what is said, than about what the speaker believes and whether it is in the hearer's interests to know what he is being told. To take all the latter information into consideration would overcomplicate the model of interaction. While I do not dispute that what a hearer believes the speaker believes etc. is important, I am not sure it is necessary to go into it for each individual utterance. Perhaps this chain of inference is hard-wired or shortcut in some way? (Leech goes on to point out that this process for the recognition of indirect speech acts is in all likelihood conventionalised over time.)

In my opinion, even in the simplest case of an assertive utterance, the hearer takes the action perspective of 'What am I supposed to do with the information the speaker is giving me?'. In this way, the onus of uptake and of contributing to the subsequent conversation is always on the hearer[8]. I digress here.

Leech does say that his list of conditions may be an oversimplification or 'idealisation' of what actually happens, but claims that it shows at least in principle that pragmatic force can be represented by heuristic procedures working on an utterance's implicatures without recourse to some conventional interpretation. He then goes on to show that the implicatures he develops match fairly closely to Searle's (1969: 65) rules for the recognition of assertions:

(*A*)	*Propositional content*:	Any proposition *P*.
(*C*)	*Preparatory*:	(1) *s* has evidence (reasons, etc.) for the truth of *P*.
(*B*) & (*D*)		(2) It is not obvious to both *s* and *h* that *h* knows (does not need to be reminded of, etc.) *P*.
	Sincerity:	*s* believes *P*.
	Essential:	Counts as an undertaking to the effect that *P* represents an actual state of affairs.

---

[7] My final model is closely related to Clark's theories on joint action. I shall be looking at a variation of his ideas in Chapter 8.

[8] This might also be the key to unifying the representation of descriptive and prescriptive utterances in my model later on; one could argue that all contributions are prescriptive to some extent in that they generate an expectation of response.

The only of Leech's conditions that does not correspond to one of Searle's is *E*, and the essential rule in Searle's classification is given no place in Leech's. Apart from this, Leech argues that, but for the propositional content condition, which is an expression of the sense of an utterance and therefore derived conventionally, all of Searle's speech act rules can be replaced by (pragmatic) implicatures.

P6: GRAMMATICAL EXPLANATIONS ARE PRIMARILY FORMAL; PRAGMATIC EXPLANATIONS ARE PRIMARILY FUNCTIONAL.

This postulate overlaps significantly with P3:

Conventional → Formal
Goal-oriented → Functional

Leech here briefly alludes to the heated debate between generative grammarians (those of the Chomskian tradition), and those who explain the interpretation of utterances in terms of their social rules and functions (such as Halliday, and discourse/conversational analysts). He argues that both approaches are necessary in order to account for the way language is used.

Functionalism is a system "which defines language as a form of communication, and therefore is concerned with showing how language works within the larger systems of human society" (op.cit.: 48). The problem with a functional explanation of language however is that we appear to have to fall back on non-empirical, probabilistic explanations of phenomena. Leech goes on to discuss the evolutionary significance of the development of language, which is of no direct relevance to my work[9].

P7: GRAMMAR IS IDEATIONAL; PRAGMATICS IS INTERPERSONAL AND TEXTUAL.

Based upon Halliday's (1973) three functions of language, Leech incorporates three different levels of functional description into his model of state transition in language understanding. These three functions are:

(1) *Interpersonal*: language functioning as a means of expressing one's attitudes and an influence upon the attitudes and behaviour of the hearer – SPEECH ACTS, ILLOCUTIONARY FORCE.

---

[9] Though he does touch upon a point that is important to my research: namely that we manage to communicate at all only because we are capable of recognising something as a symbol of something else. This observation is applicable to all levels of linguistic analysis, and indeed to any other study of human behaviour.

(2) *Ideational*: language functioning as a means of conveying and interpreting experience of the world (two further subdivisions of which are the Experiential and Logical) – PROPOSITIONS.

(3) *Textual*: language functioning as a means of constructing a text, i.e. a spoken or written instantiation of language – ACTUAL WORDS AND SOUNDS.

Leech develops these into a process model of the overall functional structure of language (shown in Figure 4.):

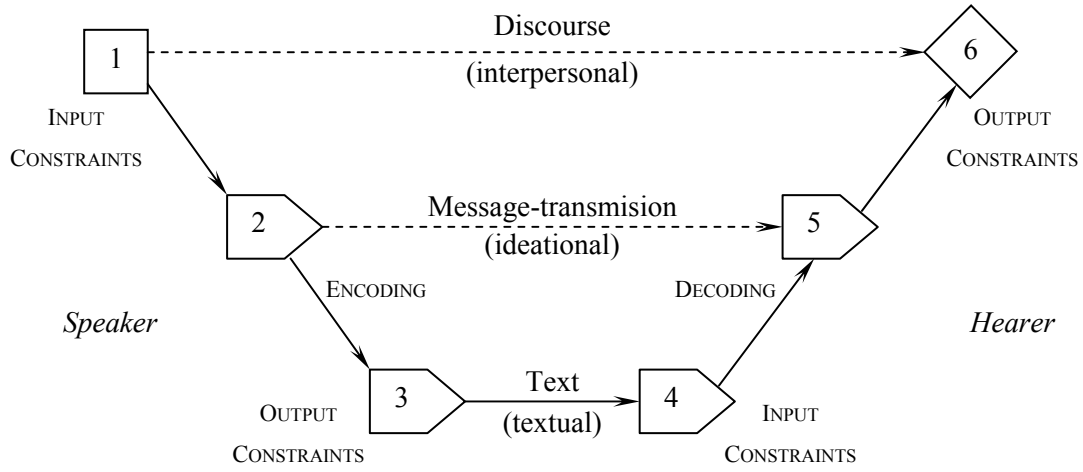


Figure 4.3 Leech's process model of language.

So, following this representation, we have:

DISCOURSE by means of MESSAGE by means of TEXT

Leech clarifies that this is not necessarily a chronological representation of consecutive processing order in time, but one of dependencies of different levels. This is an important point, for it would seem logical to me that a person has the message in his head before he thinks about the interpersonal constraints that are encoded in the text.

There are rules that govern the production of text that resemble those at the interpersonal level of communication (based on Slobin 1975):

(1) *The processibility principle*: Produce language in such a way that it is in the easiest form for others to decode. Factors such as how best to segment, highlight and order different parts of the message are relevant here.

(2) *The clarity principle*: Keep related semantic units together at the syntactic level, but also avoid ambiguity.

(3) *The economy principle*: Reduce the utterance wherever possible (while also maintaining clarity). This is I think one of the most important factors governing spoken interaction.

(4) *The expressivity principle*: Speak nicely. When it is not effectiveness that is of overriding concern in speaking, but the expression of some idea in a (culturally) pleasing aesthetic manner. The expressivity principle deals with what counts as good style in a language, and governs such situations as speech making. As such it is at a level removed from my research.

P8: IN GENERAL, GRAMMAR IS DESCRIBABLE IN TERMS OF DISCRETE AND DETERMINATE CATEGORIES; PRAGMATICS IS DESCRIBABLE IN TERMS OF CONTINUOUS AND INDETERMINATE VALUES.

This last postulate claims that grammar is much more orderly, determinate and well behaved than pragmatics. It claims that while there is some indeterminacy at the grammatical level also, this is not the general state of affairs, whereas in pragmatic interpretation it is. Leech rephrases this postulate to: GRAMMAR IS ESSENTIALLY CATEGORICAL; PRAGMATICS IS ESSENTIALLY NON-CATEGORICAL.

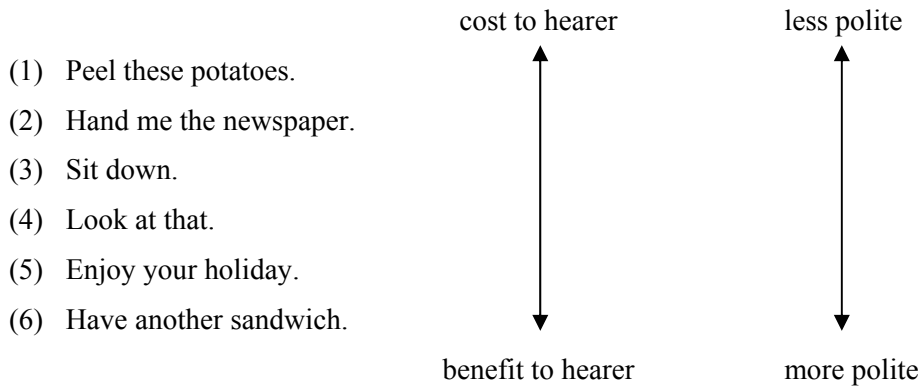
While I agree with this in broad-brush terms, it is the business of my research to try and impose a new, more rigorous structure to pragmatics, by the application of rules to its constituents: speech acts.

4.4.2 Principles of Politeness

Leech further extends Grice's principles of co-operation by the inclusion of principles of politeness in order to help explain the apparently unco-operative behaviour of people who use indirect forms to perform what I term *prescriptive* speech acts:

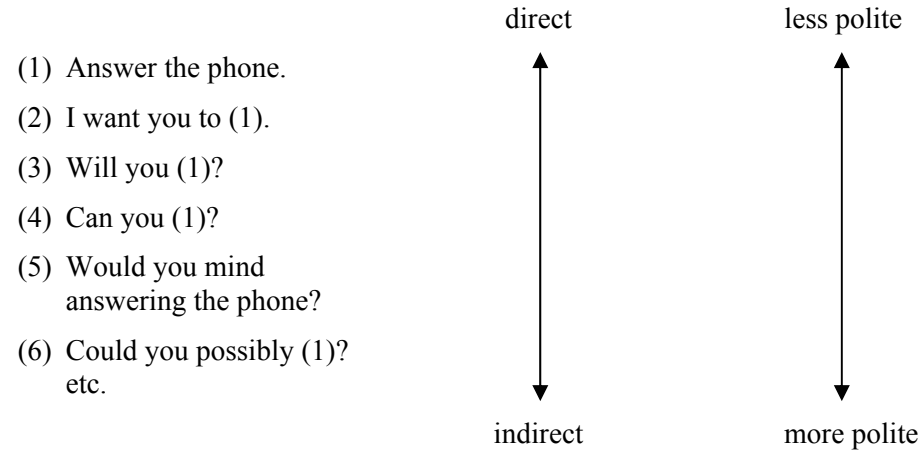
> Far from being a superficial matter of 'being civil', politeness is an important missing link between the CP [co-operative principle] and the problem of how to relate sense to force. (op. cit.: 104)

Leech defines politeness in terms of an asymmetrical scale of cost-benefit to the hearer. In other words, the more the content of an utterance will impose a cost to the hearer in terms of time and effort, the more likely it is to be constructed using a grammatical formula for politeness. Not only that, but the use of the direct imperative, which is usually considered an impolite form of address (in English), gains in politeness when complying with the action being demanded produces benefits to the agent.



Along this sliding scale, it becomes clear that some imperatives gain politeness in just the case when compliance brings increased benefit to the hearer. It should be mentioned here though that this is rather difficult to judge, and the line between cost and benefit rather dubious. Perhaps it is not so much a sliding scale, but more like a politeness *switch* between the two?

In the case when there is a cost to the hearer, there is another method of increasing politeness: that of increasing indirectness. This is perceived to be more polite because it displays (conventionally) a tacit recognition of the imposition on the hearer (and therefore an implicit openness to refusal), and also because it lessens the force of the illocution in general.



An increase in politeness here results in a proportional decrease in the utterance's adherence to the maxim of manner (see Section 4.3). So we can see how this explains why speakers often present requests in the indirect form: to maximise the politeness, at the cost of directness.

Just as Grice developed a range of maxims that give substance to the co-operative principle, Leech proposes the following set of maxims that together comprise the politeness principle (op.cit.: 132).

I. TACT MAXIM (in impositives and commissives)
(a) Minimize cost to *other* [(b) Maximize benefit to *other*]

II. GENEROSITY MAXIM (in impositives and commissives)
(a) Minimize benefit to *self* [(b) Maximize cost to *self*]

III. APPROBATION MAXIM (in expressives and assertives)
(a) Minimize dispraise of *other* [(b) Maximize praise of *other*]

IV. MODESTY MAXIM (in expressives and assertives)
(a) Minimize praise of *self* [(b) Maximize dispraise of *self*]

V. AGREEMENT MAXIM (in assertives)
(a) Minimize disagreement between self and other
[(b) Maximize agreement between self and other]

VI. SYMPATHY MAXIM (in assertives)
(a) Minimize antipathy between self and other
[(b) Maximize sympathy between self and other] (op.cit.: 132):

I have tried very briefly to summarise some of the more important aspects of Leech's work here. His principles of politeness flesh out Grice's principles of co-operative communication, and provide a significant contribution to the discussion concerning the problem of indirect speech acts. The work of both Grice and Leech has an important bearing on my own research, as both are founded on the idea that utterances are motivated actions that are anchored in context. Before turning to linguistic theories of spoken communication, I shall look at one influential piece of work in speech acts, which also tried to capture the idea of the assumption of coherency and co-operation in conversation: that of Bach and Harnish (1979). I shall also briefly sum up modern psychological theories concerning literal and inferential ascription of meaning.

4.5 Bach and Harnish's Speech Act Schema

Bach and Harnish (1979) developed an inferential approach to the analysis of speech acts. They outlined the function of a speech act schema, which attempted to demonstrate the inferences made by participants in a conversation in order to decipher whether a speaker was trying to communicate in one of four different ways:

(1) Literally[10] and Directly

(2) Literally and Indirectly

(3) Non-literally[11] and Directly

(4) Non-literally and Indirectly

In order to identify these different ways of communicating speech acts, the schema is based upon three main presumptions, and the idea of *mutual contextual beliefs* (MCBs)[12]. An MCB is the salient contextual background information necessary to decode an utterance. Bach and Harnish give the example of a woman who says 'I love you like my brother'; the MCB crucial to the interpretation of this utterance by the male addressee is that the speaker does not have amorous feelings for her brother (and therefore does not have them for him). In another case, the MCB might be different (e.g. the speaker hates her brother); in which case the interpretation of the utterance will be different also (i.e. the speaker hates the addressee). MCBs are important in determining whether an utterance is literal or non-literal, and for inferring extra information other than the content of the sentence uttered. The three presumptions that are coupled with the MCBs in the schema are as follows:

> *Linguistic Presumption* (LP): The mutual belief in the linguistic community C_L to the effect that:
>
> (i) the members of CL share L [language], and
>
> (ii) that whenever any member of *S* [speaker] utters any *e* [expression] in *L* to any other member *H* [hearer], *H* can identify what *S* is saying, given that *H* knows the meaning(s) of *e* in *L* and is aware of the appropriate background information.

---

10 Note that although Bach and Harnish do claim that there are literal interpretations of speech acts, they do not claim that there are conventional ties between the syntactic structure of a sentence and its corresponding speech act. The 'literal' acts are discarded in the schema should it become clear in the context that the speaker is not using the 'literal' form literally (i.e. Bach and Harnish's literal speech acts are defeasible).

11 An example of a non-literal speech act would be the utterance of a sentence like: 'I really love Des O'Connor's greatest hits; they're great'. The speaker in this case is, in context, obviously not being serious, so the hearer would infer a non-literal interpretation. Tone of voice often aids the identification of such non-literal speech acts. A fairly contemporary example of the recognition of this phenomenon as a means to produce humour would be the comedy sketch from the 1990s TV show *The Mary Whitehouse Experience* called 'Ray – the man cursed with a sarcastic tone of voice'. The humour of this sketch was based on the unfortunate speech disability of the main character: anything he meant to say sincerely he would say in a sarcastic tone of voice, and vice versa.

12 But note my caveat on the idea of talking about **mutual** contextual beliefs as outlined in Section 2.3.3. I assume that these are only the subset of the speaker's beliefs that the hearer needs in order to interpret the utterance correctly.

> *Communicative Presumption* (CP): The mutual belief in the linguistic community C_L to the effect that whenever a member of *S* says something in *L* to another member *H*, *S* is doing so with some recognisable illocutionary intent.
>
> *Presumption of Literalness* (PL): The mutual belief in the linguistic community C_L to the effect that whenever any member *S* utters any *e* in *L* to any other member *H*, if *S* could (under the circumstances) be speaking literally, then *S* is speaking literally. (op. cit.: 60-1)

Thus they develop the outline of a schema, as shown in Table 4.1, to deal with all four different ways of communicating listed above (adapted from op. cit.: 76-7).

	Inference	***Basis***
(1)	*S* is uttering *e*.	Hearing *S* utter *e*
(2)	*S* means... by *e*.	1, LP, MCBs
(3)	*S* is saying that *(...*p*...).	2, LP, MCBs
(4)	*S*, if speaking literally, is *F**-ing that *p*.	3, CP, MCBs
	Either (Literal Direct)	
(5)	*S* could be *F**-ing that *p*.	4, MCBs
(6)	*S* is *F**-ing that *p*.	5, PL
	And possibly (Literal Indirect)	
(7)	*S* could not be merely *F**-ing that *p*.	6, MCBs
(8)	There is some *F*-ing that *p* connected in a way identifiable under the circumstances to *F**-ing that *p*, such that in *F**-ing that *p*, *S* could also be *F*-ing that *P*.	7, CP
(9)	*S* is *F**-ing that *p* and thereby *F*-ing that *P*.	8, MCBs
	Or (Non-literal Direct)	
(5′)	*S* could not (under the circumstances) be *F**-ing that p.	4, MCBs
(6′)	Under the circumstances there is a certain recognisable relation *R* between saying that *p* and some *F*-ing that *P*, such that *S* could be *F*-ing that *P*.	3, 5′, CP
(7′)	*S* is *F*-ing that *P*.	6′, MCBs
	And possibly (Non-literal Indirect)	
(8′)	*S* could not merely be *F*-ing that *P*.	7′, MCBs
(9′)	There is some *F′*-ing that *Q* connected in a way identifiable under the circumstances to *F*-ing that *P*, such that in *F*-ing that *P*, *S* could also be *F′*-ing that *Q*.	8′, CP
(10)	*S* is *F*-ing that *P* and thereby *F′*-ing that *Q*.	9′, MCBs

Table 4.1 Bach and Harnish's schema for utterance interpretation.

KEY to Table 4.1		
...	=	what *S* means by the expression *e*
*, ´	=	dummy indicators for sentence type (declarative, imperative, interrogative)
p, q	=	proposition
*(...p...)	=	what is said is a function of the intended meaning of *e*
F	=	force of the illocutionary act
P, Q	=	propositional content of the illocutionary act

Given the schema in Table 4.1, the hearer identifies the illocutionary act by applying the following inference strategies[13] to the utterance:

Locutionary Strategy (LS): Given 1, infer 2, 3, 4.
Literal Direct Strategy (LDS): Given 4 (from LS), infer 5, 6.
Literal Indirect Strategy (LIS): Given 6 (from LDS), infer 7, 8, 9.
Non-literal Direct Strategy (NDS): Given 4 (from LS), infer 5´, 6´, 7´.
Non-literal Indirect Strategy (NIS): Given 7´ (from NDS), infer 8´, 9´, 10.

It can be seen that these are the beginnings of an extremely thorough theory of inference for the identification of speech acts. Bach and Harnish appeal to the maxims of co-operation as detailed by Grice (see Section 4.3) and use the concept of MCBs in order to infer speaker meaning. Their approach leaves at least one crucial detail unaccounted for, however. The MCBs should be altered every time an utterance is made (i.e. the context changes after every speech act), but Bach and Harnish do not discuss how this feature could be added. I shall incorporate this observation into my own model, as described in Chapters 8 and 9.

There have been a number of criticisms levelled at the inferential approach to language understanding (or what might be termed the 'standard pragmatic view') as outlined in this chapter so far. It is to these that I wish to turn briefly now.

---

[13] Note that the names of two of these strategies are modified slightly from *Literally-based Indirect Strategy* and *Nonliterally-based Indirect Strategy* in the original, for reasons of consistency and personal preference.

4.6 Inferential vs. Direct Access Interpretation

The standard inferential pragmatic view (as followed by Grice 1975, 1978, Leech[14] 1983 and Bach and Harnish 1979, above), has it that in order to analyse indirect and figurative utterances, one must first comprehend the full literal meaning, before pragmatic inferencing is called into play to infer the non-literal, underlying message.

This however causes problems when trying to analyse the fundamental, literal meaning of metaphorical utterances, such as (example taken from Gibbs 2002):

> 'Cigarettes are time bombs.'

If a hearer is to ascribe a literal meaning to such an utterance first, he would have to follow a line of reasoning that goes something like this:

(1) The utterance literally means: 'Cigarettes are kinds of bomb that are set to explode after a certain amount of time has elapsed'.

(2) This is clearly untrue (cigarettes ≠ bombs).

(3) Therefore either the speaker is lying to me, or he intends me to infer some other meaning in the context.

If the hearer does not follow this line of interpretation, he must at least infer that the speaker is producing an utterance that is not relevant and coherent in the context, before trying to find some other explanation that is consistent with the speaker acting in a co-operative and rational manner. Somehow the hearer must infer by analogy that what the speaker means is: 'The act of smoking cigarettes has a detrimental, and eventually potentially lethal effect (like that of a time bomb) to the health of a person after a certain period of time'.

The implication of the inferential analysis of indirect and figurative utterances is that literal utterances are less demanding for the hearer to understand, because he does not have to process the utterance any further in order to find out what meaning the speaker actually intended to convey. This suggests that using non-literal language should in theory always take more effort and time to process than a roughly equivalent literal utterance[15].

---

[14] Although I include Leech in my discussion of the inferential approach here, it must be noted that he accommodates the possibility of pragmatic interpretations becoming fossilised by use into conventional grammatical entities.

[15] This begs the question, why do we bother to speak in a non-literal manner at all?

So, according to the inferential view, the meaning of literal, direct utterances is determined by accessing semantic information, while the contextual implications of non-literal, indirect, or figurative utterances require pragmatic processing, which is less easy to perform.

However, recent experiments in psycholinguistics would appear to contradict this conclusion (for example, see Gibbs 1994, or Glucksberg 1998). In results for some tests of reading and phrase classification time, people seem to be able to understand (and therefore react appropriately to) figurative uses of language, such as metaphor, proverbs/idiom, irony/sarcasm and indirect speech acts just as fast as (and in some cases faster than) the literal versions, in a variety of differing contexts. This shows that it is psychologically unlikely that hearers always work out the full literal expansions of indirect and figurative uses of language, but have at least some of the structures hard-wired for quick recognition.

There is a growing school of thought that coincides with this idea, that people can understand what is traditionally considered *indirect* usage of language, in a *direct* way. Gibbs (2002) calls this the **direct access** approach.

The direct access view claims that people can interpret non-literal utterances in a direct manner when such utterances are produced in the context of realistic social settings. This is not to say that no inferences of any kind take place, just that it is not automatically the case that people work out the *complete* literal meaning *before* accessing pragmatic processes to interpret the meaning of an utterance.

Among the leading critics of direct access theories are Temple and Honeck (1999). They claim that results of equal reaction times for literal and non-literal formulations (as published by Gibbs 1986) are due to the fact that it was familiar and conventional expressions, idioms and indirect requests that were used in tests. In their own experiments, they looked at people's reaction times for non-literal interpretations of *novel* metaphors, as contrasted to that for literal interpretations of the same. The data (repeatedly) indicated that people are capable of recognising the literal meaning faster than the non-literal.

Similarly, Giora and Fein (1999) ran experiments to test people's understanding of familiar and unfamiliar sarcastic uses of a sentence in literal and non-literal settings. For example, people were asked to indicate a single word interpretation (either 'angry' or 'useful') for the use of the phrase 'Thanks for your help' coming after: (i) a description of someone being in fact helpful, and (ii) a description of someone failing completely to be helpful, and in fact proving to be a hindrance. Participants were shown the test word interpretation cues at 150ms and 1000ms after they had finished reading. The results showed that, after 150ms people responded faster to the literal interpretation than to the sarcastic one when reading unfamiliar phrases. In all tests, there was no significant reaction time difference after 1000ms. These results therefore also favour the

view that at least initially, for novel or unfamiliar formulations, we access the (salient) literal meaning first.

Empirical work in neuro-psychology also backs up this idea that our basic understanding is informed by literal (semantic) processing first. McDonald and Pearce (1996) noticed that some patients suffering from damage to their frontal lobe, as a consequence were unable to understand ambiguous language, or conventional indirect speech act use, or sarcastic remarks (where the correct interpretation of the intended meaning required the inference of the opposite meaning to that of the literal one). In other words, patients were perfectly able to understand the surface literal meaning of the sentence, but unable to make the mental leap of interpretation needed to identify the intended one in context.

Gibbs (2002) criticises the underlying methodologies of these experiments. He claims that Temple and Honeck (1999) fail to produce their novel metaphors in an appropriate context, and so the results cannot be said to be conclusive; similarly McDonald and Pierce (1996) only give the immediately preceding utterance as a context for their test sentence. He also says that Giora and Fein (1999) fail to take into account that the single word response in the case of literal interpretation is made as a trigger to a single word in the test sentence (such as 'help', in 'Thank you for your help'), whereas the sarcastic interpretation requires an analysis of the whole sentence.

There is a lot of controversy surrounding how much processing of literal meaning we perform. It is unlikely that we will ever know for sure what the underlying mental methods are for accessing information and inferring meaning; so we therefore cannot be sure that results indicating differences in reading times for literal and non-literal sentences (in the different contexts that set up the different interpretations) are really to do with the difference in interpretation time of the sentence being tested, or due to some other factors that might affect processing (such as for example the form of the preceding context). As it may not be possible to test consistently two uses of a sentence under exactly the same conditions, results cannot be relied upon to show conclusively one way or another whether people use semantic or pragmatic processes to interpret indirect forms in context.

One of the endemic problems for this debate between inferential and direct access interpretation is that there is no consensus as to what 'literal meaning' actually means. Do we access the literal meanings of the words themselves? If so, then that means that in a sentence such as 'I can't stand to stay here any longer', we must analyse the word 'stand' (and indeed 'stay') to mean a variety of different things. The dictionary definition of 'stand' gives a long list of alternative meanings: which one is the literal one? Do all these meanings of the word 'stand' rush through our minds when we consider the utterance of this sentence?

I would suggest not. If indeed we do resort to an analysis of literal meaning at the word level when understanding language, then this literal meaning must be a restricted one defined by the immediate sentential context and also the context of the conversation and situation as a whole.

This could explain why it may be possible that even novel uses of figurative and indirect language can have imperceptible reaction time differences when utterances are produced in context. Thus the context itself might set up a non-literal interpretation of an utterance. Gibbs (2002: 462) says:

> People may still need to draw complex inferences when understanding some ironic statements, but part of these inferences can occur before one actually encounters an ironic utterance.

In fact, possibly we are in general 'set up' to recognise the performance of non-literal utterance because as part of our cultural background, we hold some expectation that speakers will not always be uttering something literally. This would account for the variation in the use of literal and non-literal configurations of utterances in different languages and cultures[16].

I shall not dwell for much longer on the debate between these opposed theoretical positions – after all, any conclusion can only be based on little more than conjecture. My own view is that we learn how to deal with figurative and indirect utterances as we are growing up. I would suggest that our mental models of word and phrase use are, from an early age, primed for multiple possible interpretations, and that repeated use of a word or set of words to stand for a concept, will reinforce the interpretation and thereby conventionalise its use for us in future interactions. I would argue that we probably perform some deep level inferencing at least for the first time that we hear a new figurative use of language, so that we can apply the generalised interpretation of the specific case when next we hear the same figurative use again. So, for example, the proverb:

'Out of the frying pan and into the fire.'

might be interpreted as some generalised semantic framework of meaning, such as:

'To go from a bad situation into a worse situation.'

---

[16] It is interesting to note in some cultures, non-literalness is built into the language structure itself. In (Argentine) Spanish, non-literalness is used to distance the speaker from taking responsibility for the occurrence of an accident that will perhaps anger the hearer, or cause the speaker to be blameworthy in some way. For instance, when reporting the breaking of a favourite vase, a speaker is very likely to say 'se rompió' ('it broke itself') rather than the more factually correct 'lo rompí' ('I broke it'). The non-literalness of 'se rompió' is not noticed by a native speaker – it is heard and interpreted as 'lo rompí'.

In this way, the next time we hear the proverb, we can apply the generalised semantic representation to the current situation and so shortcut the inferencing process. The same kind of analysis would work for metaphors.

There are however some turns of phrase that I simply cannot accept that we ever analyse literally at all. Take for example the English phrase 'to kick the bucket'. I suggest that we have a direct link to the interpretation of this phrase as meaning something roughly the same as the verb 'to die', while also having connotations of informality. It is unlikely in the extreme that we ever literally think of someone physically kicking a bucket before realising that this is inconsistent in the context. Again, I suggest that we shortcut the inferencing process by assuming the phrase as part of our lexicon.

So, while I am in agreement with some of the ideas suggested by the direct access approach to understanding utterances, I do not think that we can do away with the inferential approach. We clearly need to process and apply new concepts to some extent. If we look at humorous utterances, I think this becomes clearer. It can sometimes take a hearer quite a while for the 'penny to drop' when trying to understand a joke; and in some cases, the hearer might have to admit that he just does not 'get it'. Anyone who has had anything to do with children will know that for a considerable stage in their development, children do not understand jokes and need to have them explained to them – not only what the meanings are, but also why they are funny. I believe that this shows the beginnings of our analysis and acquisition of non-literal interpretations.

In summary then, I would agree with a weak form of direct access theory; there is strong evidence (Lakoff 1987, Glucksberg 1991, Gibbs 1994, Goldberg 1995, Narayanan 1997, Chang and Fischer 2000) that we store semantic representations of meaning at a higher level than that of the word, or even phrase. Thus idioms, proverbs and metaphors are generalised[17] for quick application in context, and indirect speech acts are conventionalised for reasons of politeness, and also to circumvent unnecessary conversational steps. According to my view therefore, inferential pragmatics has the last word in utterance interpretation.

I do not pretend to have done justice to all the arguments concerning literal meaning here; this in itself might form the substance for another dissertation. But I have tried to position my own research, as well as that of other people's, with respect to what is in essence a semantic/pragmatic divide. I myself feel that this divide is an artificial one, and that some

---

[17] This is borne out by the existence of different versions of the same generalised idea contained in a proverb that occur across different languages. For example, in Spanish, the proverb 'The straw that broke the camel's back' is realised by 'La gota que rebalsa el vaso' (which means 'The drop that causes the glass to overflow').

middle ground needs to be found to accommodate both positions (perhaps a complementarist view, as defined by Leech).

4.7 Summary

In this chapter (and previous ones) I have attempted to establish the following:

(1) Typically, utterances not only express propositions (i.e. their 'content'), but also, together with their illocutionary force, perform actions.

(2) The complete characterisation of speech acts relies upon the fulfilment of preconditions.

(3) Explicit illocutionary acts have the force explicitly named by the illocutionary verb in the matrix clause (although the content of such acts need not always be taken literally: e.g. note the difference between 'I warn you that there is a rabid dog in the kitchen' and 'I warn you that there is a marshmallow in the kitchen').

(4) The three major sentence-types in English (declarative, imperative, and interrogative) need not always have the forces traditionally associated with them (stating, ordering/requesting, and questioning respectively); explicit illocutionary acts are always in declarative form.

(5) Illocutionary force cannot be expressed in truth-conditional semantics.

(6) Illocutionary force indicates what is to be done with the propositional content of an utterance (e.g. an assertion of a proposition is intended to be believed by the addressee; with a command the addressee is meant to carry out the proposition).

(7) In a conversation, it is assumed that there is some underlying principle of co-operation between participants (i.e. that in speaking, participants are following certain maxims and are attempting to convey some relevant, coherent piece of information, for some identifiable reason or intent).

(8) Utterances not only perform speech acts, but by the identification of these speech acts the participants' contextual beliefs are altered (e.g. if I state that *X*, the fact that I am committed to *X* in some way will have been added to the context and will play a part in the subsequent conversation. This gives rise to the idea of speaker *commitments* – see Chapter 9 for details).

(9) It is unclear where one can draw a line between semantic representation and pragmatic inference in determining and ascribing meaning to an entire utterance in context.

(10) Utterances rarely, if ever, have one and only one interpretation, even when bounded by the constraints of physical, mental and conversational contexts.

These assumptions treat illocutionary acts in a pragmatic manner; a theory that would be able to handle such acts would need to take these assumptions into consideration, and work accordingly in a thoroughly pragmatic way. Before outlining the approach that I have adopted (and adapted) to implement a new model for inferencing, I shall first consider some of the linguistic approaches to discourse analysis, as well as other computational models of speech acts that have been developed.

Chapter 5

Linguistic Approaches to Discourse Analysis

A.	Hvad er egentlig en såkaldt vittighed?	*What is a so-called joke really?*
B.	Jo, ser du… A. siger noget til B. – og så siger B. noget til A. – og så ler man ad det!	*Well, you see…* *A. says something to B. –* *and then B. says something to A. –* *and then one laughs at it!*

(*Udvalgte Fluer*, Storm P.)

There are a number of different (what I call) linguistic approaches to studying spoken language, as shown in Figure 5.1 (which is adapted from Eggins and Slade 1997: 24). In reality these also belong to other disciplines, such as philosophy, ethnography, socio-linguistics and psychology.

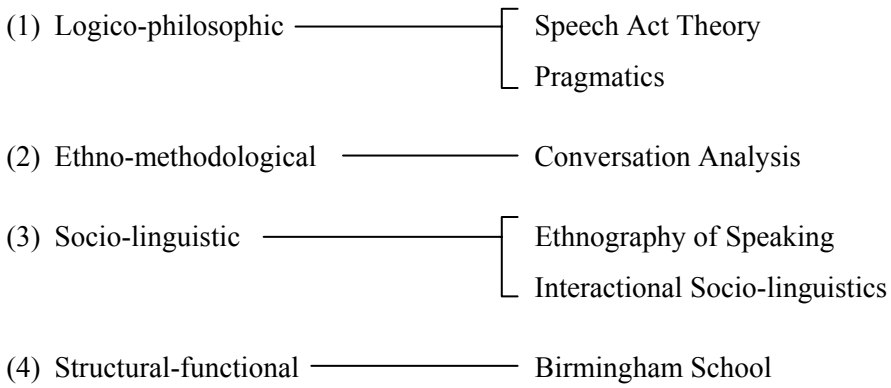


Figure 5.1 Relevant approaches to analysing general conversation.

This list is not intended to be comprehensive by any means. I will only be looking briefly at those linguistic theories that most bear on my own research. I have already dwelt on (1) at some considerable length in the first chapters of the dissertation, so this chapter will concentrate briefly on (2) – (4) above.

5.1 Conversation Analysis

One of the major concerns of conversation analysis in its infancy was how different speakers managed to take their turns appropriately in conversation. Sacks et al. (1974) suggest that speakers are able to recognise the right moment to take over in the conversation because speakers talk in **Turn Construction Units** (**TCU**s). These are grammatically complete units of language such as a sentence, clause, or phrase, which signal to the other speakers a potential gap in which to take over. This transfer of the 'floor' can be gauged by participants in a conversation in a number of different ways, some of which will occur at the same time: falling

intonation, the end of a grammatical unit of conversation, or utterance, posture and gaze or eye-contact, or even a direct invitation to one of the other participants to continue. The system of turn-allocation (derived originally from Sacks et al. 1974) can be shown in the following way:

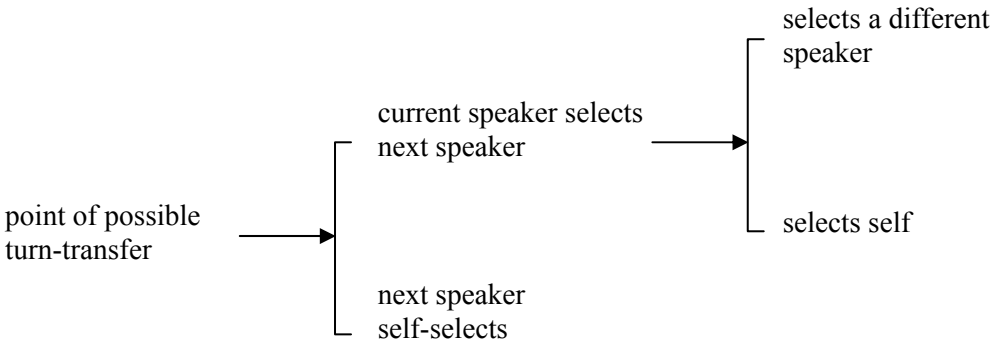


Figure 5.2 The turn taking system.

Note that it does not follow that the nominated next speaker will necessarily continue the conversation. Turn taking is not pre-negotiated, but dynamically decided in the conversation. The turn may be usurped by one of the other participants, especially in the case when there is a feeling that he is in a better position to answer or contribute to what is being said. (There are some situations where this would be inappropriate, e.g. in the context of a school classroom when turn is strictly allocated by the teacher.)

Conversation is thus characterised as a perpetual mutual effort between participants to avoid any unseemly **lapse** in the flow of talk. So, how do conversations ever stop? Sacks et al. (1974) noticed that utterances often come in opener/response pairs, called **adjacency pairs**. These pairs typically have the following in common:

(a) two-utterance length (they come in twos);
(b) adjacent positioning of component utterances (as suggested by the name);
(c) different speakers producing each utterance. (Schegloff and Sacks 1973)

The quintessential adjacency pair is the question/answer sequence. Others include: complaint/denial, compliment/rejection, challenge/rejection, request/grant, offer/accept, offer/reject and instruct/receipt, etc. (Sacks et al. 1974: 717).

It is the existence and recognition of the first part of an adjacency pair that leads to the expectation of the second part. This furnishes speakers with strong clues as to when a possible turn-transfer will take place, and indeed be expected (as well as guiding the choice of the next turn type). The second act of the adjacency pair can have a positive or a negative force. Speakers display a marked preference for positive, co-operative compliance; rejection or, non-compliance is dispreferred. The former is usually represented by shorter turns, whereas the

latter by longer ones, as speakers will feel the need to explain or justify their non-compliance. Adjacency can also be defined as **sequential relevance** to the utterance immediately prior to it, unless explicitly indicated otherwise.

Taylor and Cameron (1987: 103) bundle up the occurrence of adjacency pairs in a behavioural package:

> My behaviour is designed in light of what I expect your reaction to it will be: i.e. you will react to it as conforming to the relevant rule or as in violation of it, thereby leading you to draw certain conclusions as to why I violated the rule… Thus, by the inexorable fact that interactions progress, any component action inevitably is temporally situated in a sequential context, a context to which it is an addition and within which it will be interpreted, held accountable and responded to in turn.

In my view, one of the strengths of the contributions of conversation analysis is the reaction against intuitive methods of working with conversation and the reliance on actual recordings of 'naturally' occurring interactions, which have been carefully transcribed, often in meticulous detail. So that one can say that the observations made by conversation analysts do not come from made up, simulated dialogues, or from the setting up of artificial (often highly delimited) interactive contexts, but have come from looking at everyday exchanges to inform their observations and conclusions.

However, there are equally some major criticisms to be levelled at this work too.

(1) *Lack of systematicity*: There has been no attempt by conversation analysts to provide an exhaustive list of all possible adjacency pairs, nor indeed is there any indication about how the (first of an) adjacency pair might be recognised in the first instance. Although TCUs have been identified as the means by which people break up the conversation in order to know when to join in, it is not specified how they know when and how to identify the end of a TCU in any situation (the suggested method of recognising such boundaries is by turning to linguistics to provide the information needed). This lack of systematicity prevents quantitive, statistical analysis of the conversational analysis.

(2) *Fragmentary focus*: Analysis has mainly been carried out on small segments of conversation and not on sustained lengths of communication. The conversations themselves, while being genuine recordings of spoken interaction, are not often instances of casual conversation, so are the same conclusions applicable?

(3) *Mechanistic view of interaction*: While it explains the how, it has nothing to say about the why, nor how the application affects later interaction.

While there are many drawbacks to this approach, there are also some key concepts that I will wish to revisit when analysing the development of my own model of conversation. I wish to

show that the adjacency pair hypothesis is actually the observation of a deep underlying structure in conversation. In other words, I hope to explain the occurrence of this phenomenon from the workings of a more general model of interaction.

5.2 Conversation and Cultural Context (Ethnography of Speaking)

Ethnographers of communication are primarily concerned with the analysis of patterns of communicative behaviour, by observing how participants use language. Their aim is to try to discover how members of a specific culture perceive their experiences and then transmit their interpretations of these.

Hymes (1972) developed a schema for breaking down the constituents of a context into units of analysis, which he called the **speech events** in which the language happens. He listed these in a classificatory grid, called a SPEAKING grid (because of the letters used to identify the different components).

S	Setting, scene	Temporal and physical circumstances, subjective definition of an occasion
P	Participant	Speaker/sender/addressor/hearer/receiver/audience/ addressee
E	Ends	Purposes and goals, outcomes
A	Act sequence	Message form and content
K	Key	Tone, manner
I	Instrumentalities	Channel (verbal, non-verbal, physical forms of speech drawn from community repertoire)
N	Norms of interaction and interpretation	Specific properties attached to speaking, interpretation of norms within cultural belief system
G	Genre	Textual categories

Table 5.1 Hymes's SPEAKING grid.

This grid can be used to analyse a local cultural taxonomy of what Hymes calls 'units' of communication. These units are organised in the following structure (shown in Figure 5.3):

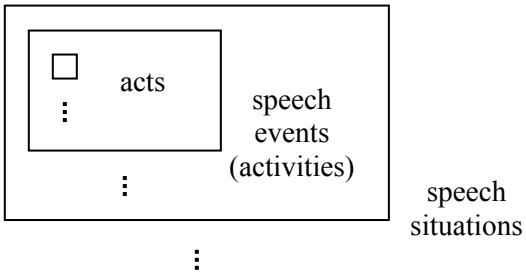


Figure 5.3 Hymes's structure of communication.

The structure he describes is very similar in some respects to that proposed by Sinclair and Coulthard (1975), which I shall be dealing with in Section 5.4.

The largest unit is the **speech situation** (the social setting in which the conversation takes place, say a meeting); the next size of unit is the speech event or activity (say, the conversation or debate that is taking place within the meeting); and the smallest unit is the speech act (say, a disagreement within the discussion within the meeting). Although Hymes does not explicitly equate the smallest units of his analysis of conversation to Searle's illocutionary acts, some of the examples he gives merit the classification of speech act, although others, like 'greeting' (Searle counted these types of act as Expressives), do not, and in fact appear to belong to a larger unit than the speech act, which is unaccounted for by Hymes.

Hymes recognises that, although all units of communication are important for the interpretation of an utterance, it is the smallest unit, the speech act, that handles the basic management of a conversation:

> Discourse may be viewed in terms of acts both syntagmatically and paradigmatically: i.e., both as a sequence of speech acts and in terms of classes of speech acts among which choice has been made at given points. (1972: 57)

Although Hymes's approach seems a very vague and cluttered way of representing a speech situation, it is significant because it was one of the first to try to include some details of the context of the speech situation to bear on studies of language understanding. Hymes's work is not of direct relevance to me, as it is predominantly concerned with cultural variations in language.

5.3 Interactional Socio-linguistics (Discourse Analysis)

This approach grew out of work by Gumperz (1982) who was in turn influenced by the work of the sociologist Goffman (1967). Like Hymes's work in ethnography of communication, they were interested in how context affects the understanding and interpretation of discourse.

5.3.1 Cultural Background

Gumperz made a detailed study of the grammar and intonation of interactions between speakers of different races (specifically Indian and native British speakers of English in England). The results of his work show that people from different social and cultural backgrounds pick out different features in discourse and interpret the tenor of the whole message differently in accordance with their perceptions of the *contextual cues* available in the discourse. So, for example, some participants in a conversation might perceive an utterance as rude and aggressive, while others think the same pattern of intonation denotes deference and consideration. This explains why cross-cultural communication is often hazardous and prone to severe misunderstandings. Gumperz suggests that this is because the interpretation of language is inseparable from the socio-cultural background in which it takes place:

> What we perceive and retain in our minds is a function of our culturally determined predisposition to perceive and assimilate. (1982:4)

Gumperz claims that due to our cultural conditioning, we can only successfully interact with those members of society with whom we share the same contextualisation cues. Gumperz's work aimed at determining and classifying the types of contextualisation cues that operate in different cultures, with a view to being able to predict sources of potential miscommunication between them.

5.3.2 Style in Conversation

Tannen's (1984) work on the characterisation and identification of different 'styles' of conversation, follows on from that of Hymes (1972) and Gumperz (1982). She is interested in how we establish and develop a *rapport* with each other in conversation. She concentrates on looking at the overall characteristics and stylistic devices we use in order to gain mutual understanding in spoken conversation, rather than placing importance on the sequential organisation that we use in order to do this.

Tannen identifies her strategies for the creation of rapport between speakers, with Lakoff's 'Rules of Rapport':

1. Don't impose (Distance).
2. Give options (Deference).
3. Be friendly (Camaraderie). (Lakoff in Tannen 1984:11)

In essence, Tannen creates a very specialised cultural set of principles that govern conversation in much the same way as do the principles developed by Grice and Leech (discussed in Chapter 4). She analyses conversation according to a number of different features, and the rules that

govern them, among which are: topic choice and management, pacing of conversation, tactics for co-operative completion and feedback, narrative strategies and use of pitch, intonation and quality of voice. She specifies various speaker strategies in terms of the preferences and conventions they use to convey certain impressions by the manipulation of the features listed above. For example, Tannen gives the preferences for 'narrative strategies' as follows:

1. Tell more stories.
2. Tell stories in rounds.
3. Prefer internal evaluation (i.e. point of a story is dramatized rather than lexicalised). (1984: 30-1)

Tannen's analysis of conversation as a variety of interactive styles as determined by the social group activity in which we are currently participating highlights the extent to which language use is anchored to social and cultural context. She claims that it is the use of these different interactive styles that aligns and identifies us with one particular group or another.

Tannen also sees these varying styles as indicating differing but equal modes of communication. Cameron (1992) and Eggins and Slade (1997) however reject this claim and argue that there are some conversational styles and strategies that can seriously disadvantage and subordinate a speaker within a conversation. While Tannen's work is useful for showing how important variation is in conversational behaviour, not all her observations can be applied generically to all types of conversation.

5.3.3 Markers for Conversational Coherence and Continuity

My namesake, (Deborah) Schiffrin[1] (1987) carried out a close study of a number of words and phrases, which she terms *discourse markers*. Unlike her colleague Tannen, who tries to define more general features of conversation, Schiffrin focuses on the function of various discourse markers (such as *oh, well, but, or, so, because, now, then, y'know* and *I mean*) which indicate to the hearer the beginning or end of a unit of talk, and position the speaker with respect to the content of their utterance. The analysis she provides is based on the assumption that talk is organised by the speaker so that the meaning is easily accessible to the hearer. The discourse markers that she investigates function to give cues to the hearer as to the status of the current utterance and throw up expectations of the kind of interpretation being elicited. Schiffrin looks at a restricted set of discourse markers and their distribution and function within conversation (which is taken from peer-group recordings of interviews, not from naturally occurring general

---

[1] Out of interest, as far as we are aware or can ascertain, Deborah and I are not directly related, although we can both trace our ancestry back to Russia.

conversation). Although this set is quite limited in number, the conclusions she draws from the observation of their roles in spoken interaction have important implications for the way we structure our talk, and offer insights into how the use of certain words indicate sequential relevance in conversation.

For example, Schiffrin's analysis of the discourse marker *oh* leads her to conclude that it is used as a marker of information management. It marks a transition in the speaker's information state, and signals this internal cognitive process to the hearer.

Schiffrin's work has important significance for my own research. The discourse markers she studies function as information managers at the pragmatic level. If we look at the way that the discourse connectives *and, but,* and *or* are represented in her analysis the relevance of her work here becomes clearer (taken from Schiffrin 1987:190).

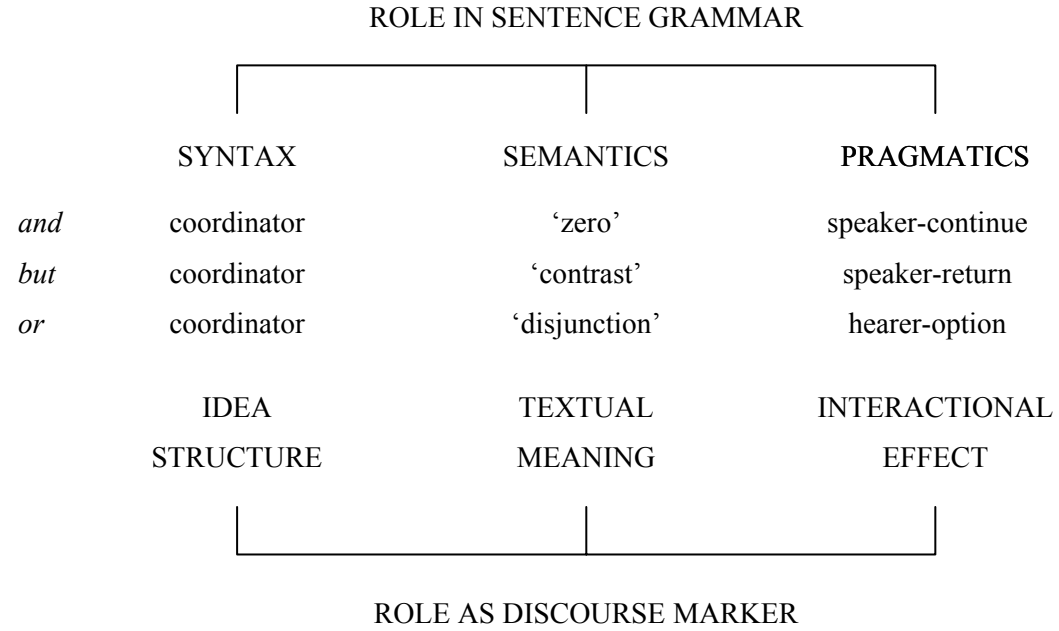


Figure 5.4 Conjunctions as discourse markers.

Similarly for the phrases *y'know* and *I mean* which mark levels of participation and inclusion for the hearer.

One of the problems with classifying discourse markers in this way is that these really are very language specific. Although they perform pragmatic functions in conversation, they are conventionalised and belong to the lexicon of the English language. Therefore the analysis developed by Schiffrin is inapplicable to any other language – a separate study would have to be carried out in order to determine whether there were comparable discourse markers in other languages. This brings us back to the question discussed in Chapter 4 about whether it is ever possible to separate the pragmatic features of a language from its lexicon.

I shall now briefly consider a structural approach of a more global level to analysing interaction as epitomised by the Birmingham School.

5.4 The Birmingham School

This comes under the structural-functional approach heading in Figure 5.2 at the start of this chapter. This approach first stemmed from work carried out in socio-semantic linguistic theory as first set out by Firth (1957) and later in Palmer (1968) and Halliday (1961). However the Birmingham School evolved towards work on discourse structure, whereas Halliday concentrated on the systemic and semiotic aspects.

The contribution of the Birmingham School (as typified by Sinclair and Coulthard 1975) was to recognise that discourse was a level of language organisation in itself, quite distinct from grammar and phonology. Discourse units were defined for the study of interactive talk (especially developed in the analysis of classroom discourse). They saw language as made up of the following levels:

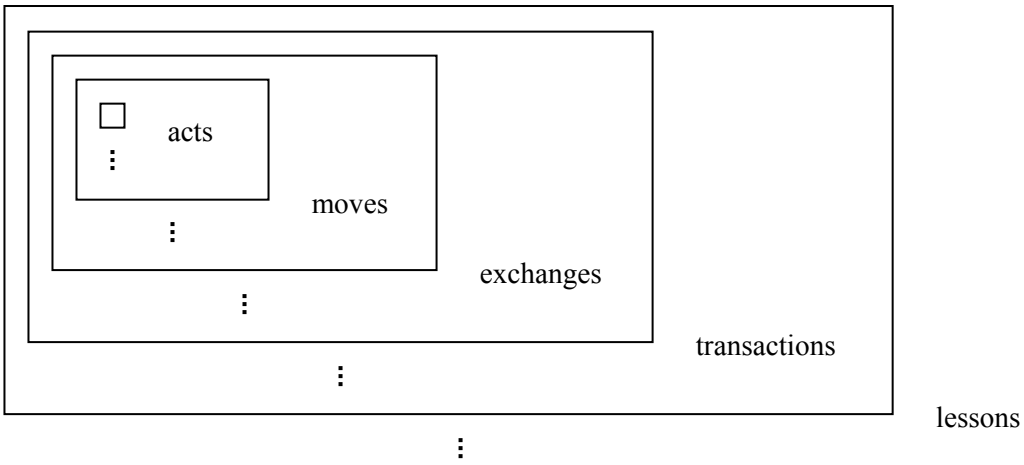


Figure 5.5 The structure of Sinclair and Coulthard's classroom interaction.

Lessons are the largest unit of discourse in a pedagogical setting. The aim was to define the links between these discourse units and grammatical units such as the clause.

The Birmingham School is best known for the work carried out on identifying the structure of conversational exchanges. According to this approach, the exchange is the fundamental unit of language in general conversation, which contains the sequence of turns in a conversation whose analysis is determined according to certain functionally expected moves. This was an improvement on the approach of conversational analysts, which only looked at adjacency pair two-turn sequences, and did not attempt to expand their theories to a more general structure of discourse.

Sinclair and Coulthard (1975: 21) proposed that "two or more utterances" go to make up a pedagogic exchange:

$$\text{Initiation} \wedge (\text{Response}) \wedge (\text{Feedback})$$

So a typical example might be:

Teacher: What is the capital city of Argentina?	*Initiation*
Student: Buenos Aires.	*Response*
Teacher: That's right.	*Feedback*

There are some obvious problems with this description of interaction that spring immediately to mind as soon as one tries to apply this model to any other conversational exchanges. The model reflects the particular structure of classroom interaction; general conversation differs in a variety of ways.

(1) *Informal vs. formal*: The teacher has control of the order of turns and students are only allowed to take turns under strict supervision. Unlike in general conversation[2], there is a power imbalance.

(2) *Completion vs. continuation*: In pedagogic exchanges the speakers look for a quick way of closing down the current topic, whereas conversation tends towards open-endedness. There is no rush to get to the end.

(3) *Direction of benefit*: It is not usually the case that someone will ask a question to which they already know the answer (unless they are looking for some kind of affirmation from their audience for a belief they hold). The point of a question in a classroom setting is to test the knowledge of the student and not because the teacher really wants to be informed. While these kinds of exchanges can occur in general chat (as when for example a participant wishes to lead one of his hearers to a logical deduction by means of answering questions that lead to it), they do not do so as a rule. The direction of benefit therefore is generally different in conversation; whereas in a class the answer benefits the responding participant, in general conversation it benefits the initiating participant:

		class		**general**
Speaker1:	Question		***benefit***	
Speaker2:	Answer			

[2] Note that this is a bit of an over-generalisation. Even in the most casual conversation there may be a power imbalance due to age, gender, relationship, etc.

Therefore, as we can see, the purpose of the response is different in the context of general conversation. This also explains why the questioner is less likely to then produce feedback in the same way – one does not need to affirm that a response is correct when one did not know the answer.

(4) *Dialogue vs. multilogue*: This model takes no notice of the potential for other speakers to contribute. The exchange pattern is strictly observed between two people. In casual conversation this is not the case; more than two people can comfortably take part in exchanges.

Attempts were made to come up with a generalisation of Sinclair and Coulthard's (1975) model with varying degrees of success. Coulthard and Brazil (1979) changed the third item 'feedback' for a less specific 'follow-up' and also suggested that Responses can play a dual role of both response and new initiation:

$$\text{Initiation} \wedge (\text{Re-Initiation}) \wedge \text{Response} \wedge (\text{Follow-up})$$

Exchanges now consist of two to four utterances. They further note that:

(1) Follow-up itself could be followed up.

(2) Opening moves indicate the start of the exchange sometimes, which do not restrict the type of the next move.

(3) Closing moves sometimes occur which are not necessarily a follow-up.

When these observations are added to their formula we end up with:

$$(\text{Open}) \wedge \text{Initiation} \wedge (\text{Re-Initiation}) \wedge \text{Response} \wedge (\text{Feedback}) \wedge (\text{Follow-up}) \wedge (\text{Close})$$

This now can deal with anything from two to seven move exchanges. However, when applied to a conversation, we see that this still does not quite work. The model is too rigid and does not allow for the exchange to be broken by another exchange, or recognise that often conversations have a very nested structure. Furthermore, identifying the boundaries for exchanges is very difficult – there is no method for doing this included in their account. Indeed, even how to identify individual moves is not made clear.

Later efforts to fit this structure to general conversation by Berry (1981) arguably generalised the occurrence of the Response move to such an extent that the functional description became meaningless. It appears to me that researchers in this particular field lost sight of the objective of a functional explanation.

Stenström (1994) and Tsui (1994) also generalise the ideas of Sinclair and Coulthard (1975), along with other conversation and discourse analysts, to spoken interaction. Stenström describes the systematic over-all move structure for interactional exchanges, which I have represented diagrammatically in Figure 5.6.

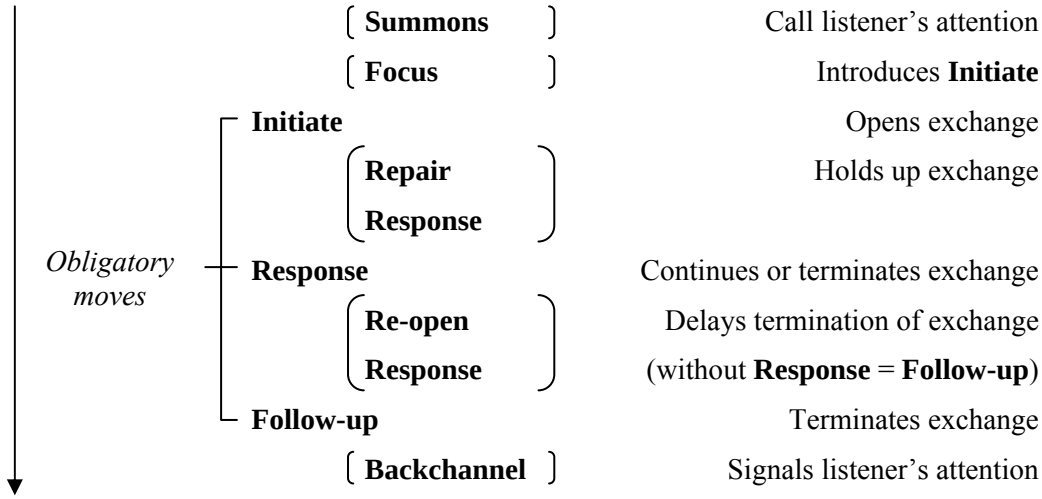


Figure 5.6 Stenström's interactional move structure

Again there is no allowance for exchanges to be broken up and nested within the conversation as a whole (that is not necessarily to say that this structural description will not be extendible to account for this phenomenon). Some of the moves are not clearly compartmentalised. There appear sometimes to be cross-type moves; for example, a Re-open without a Response is counted as a Follow-up (as I indicate in Figure 5.6). Often it is difficult to decide which type an utterance should be; the difference between Backchannels and positive Responses, for instance.

Elements of structure	**I**	**R**	**F_1**	**F_2**
Move	Initiating	Responding	Follow-up (1)	Follow-up (2)
Head act: primary class	Initiating (Initiation)	Responding (Response)	Follow-up (1)	Follow-up (2)
Head act: subclass	Elicitation Requestive Directive Informative	Positive Negative Temporisation	Endorsement Concession Acknowledgement	Turn-passing

Table 5.2 Tsui's taxonomy of discourse acts

Tsui (1994) outlines an extremely thorough approach to the characterisation of discourse acts according to a three-part transaction. She argues against Burton (1981) who said that three-part exchanges are highly classroom specific, claiming that even in general conversation, "the

addressee… displays his or her interpretation in the response" (ibid.: 32). These three part transactions are shown in Table 5.2.

The systems of choice for Initiating act are shown in Figure 5.7, for Responding in Figure 5.8 and for Follow-up in Figure 5.9.

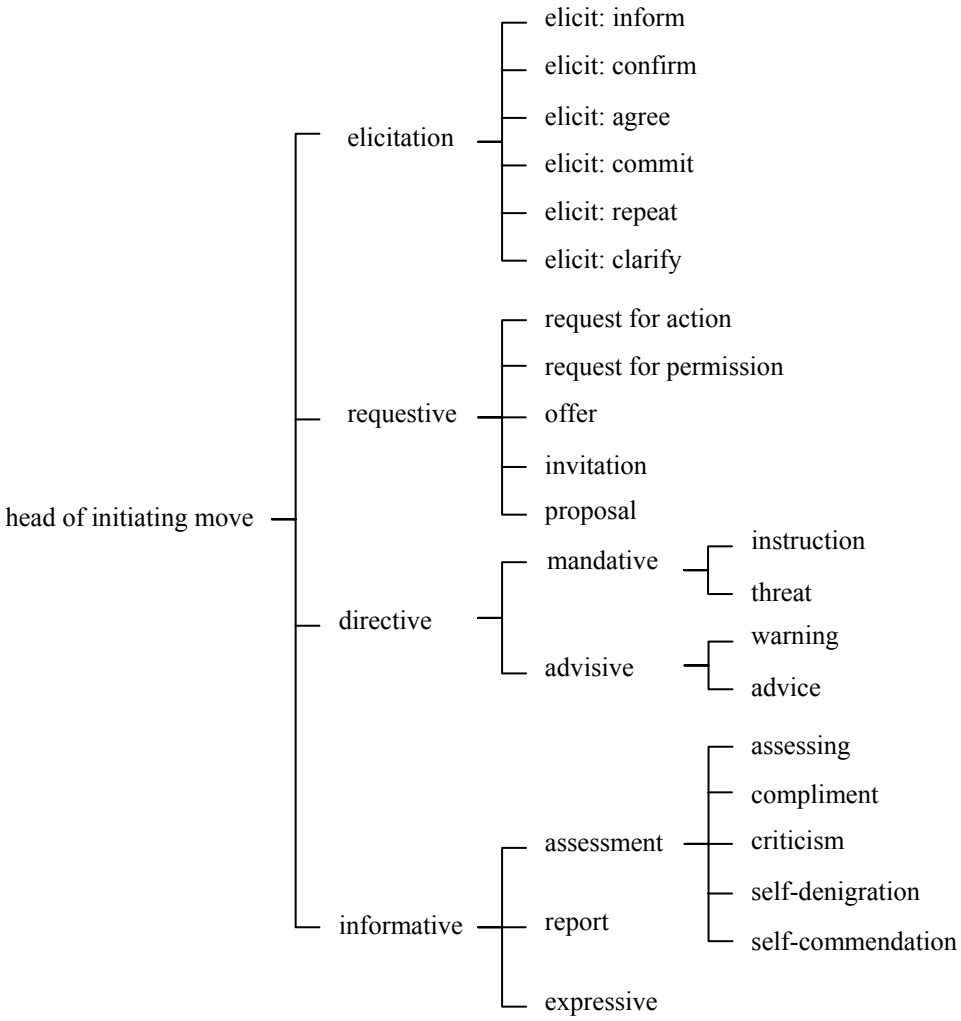


Figure 5.7 Tsui's systems of choices at the head of initiating move.

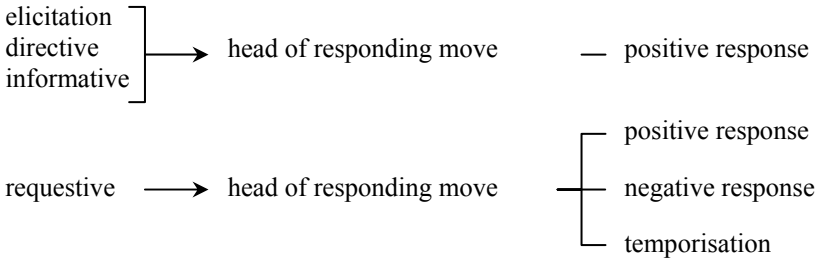


Figure 5.8 Tsui's systems of choices at the head of responding move.

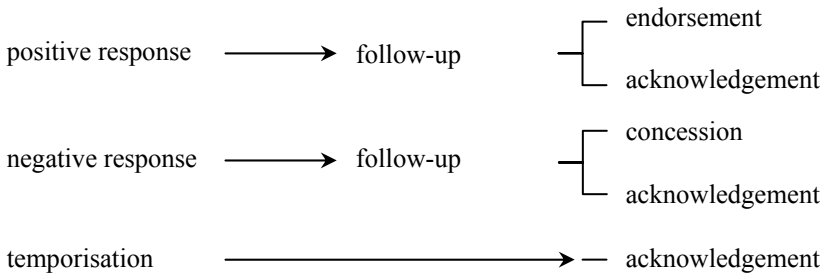


Figure 5.8 Tsui's systems of choices at follow-up.

Of all the linguistic approaches that I have reviewed so far, Tsui's structural treatment of conversation is the closest relative to the model that I have developed. Her classification of discourse acts by whether they are initiating, responding, first follow-up or second follow-up, is similar to the system I develop, which will consider what has gone before in the assignment of speech function. I disagree with some of the theoretical distinctions that she draws, especially in her initiating acts, but I will address these when I discuss the development of my own model. In fact I shall be returning to both Stenström and Tsui in Chapter 9 when I discuss questions and requests.

Having spent some time reviewing some of the linguistic attempts to define the structure of discourse, I now look at the various computational models of speech acts in dialogue. The aim is to highlight differences in methodological outlook, and explain why the computational models currently developed are inadequate for dealing with unrestricted general conversation.

Chapter 6

Computational Models of Speech Acts

> 'I've got a plan so cunning you could put a tail on it and call it a weasel.'
> (*Blackadder III*, Richard Curtis and Ben Elton)

As little as twenty years ago, it would have been difficult to find other researchers working with computational models of speech acts; indeed one would have been hard pressed to find relevant background even in a similar field. During the 1990s however, interest in using pragmatic features of conversation to aid language understanding in computers expanded phenomenally. Nowadays it's not a question of finding something to include in a survey of the state of the art in speech act modelling, but of choosing what to exclude. I make little apology therefore for the references to potentially relevant work that I might have left out from the discussion in this chapter. I have tried to limit myself to those key research projects that bear directly on my own research, and those that typify the two main approaches to the problem of identifying or modelling speech acts.

The first approach is based on the logico-philosophical theories of the likes of Searle and Grice (as covered in Chapters 3 and 4), and is an approach that tends to ascribe meaning in terms of the speaker's *plans* and *intentions*. The mental attitudes of co-operating agents are modelled in order to be able to reason about each other's goals and make plans to carry out their own. This approach is characterised by the modelling of dynamic factors of mental state (one's own and other agents), such as beliefs about what agents know about the world, desires about how the agents wish the world to be changed (their goals), as well as intentions which are plans concerning how to bring these changes about.

The second approach is based on linguistic analyses, such as those of Sacks and Schegloff, and Sinclair and Coulthard (as covered in Chapter 5), who view the ascription of meaning according to expectations thrown up by conversational *structure* and lexical or phrasal *cues* as perceived by the hearer. This approach is founded on the idea that there is a grammar of discourse, which is modelled using finite-state models or state transition networks in order to map out the course of 'legal' dialogues. The idea is that the performance of each utterance constrains the choices for subsequent utterances. This approach is particularly popular in automated dialogue systems today and it is this approach that I shall be taking in the development of my model in Chapter 9.

Of course I shall be arguing that in the final analysis both approaches are necessary in modelling conversation accurately[1].

6.1 Plan- and Intention-Based Analyses

There have been several attempts to formalise speech act theory for use in *natural language processing* (NLP) computer programs. These attempts have had varying degrees of success, and have helped to clarify the problems that accompany the computational schematisation of speech act theory.

6.1.1 Plan Recognition

One of the most influential pieces of work in this area was carried out by Allen (1983, 1987), who based his program on earlier work by his colleagues Cohen and Perrault (1979). Allen was interested in trying to identify and predict helpful linguistic behaviour in conversation. He analysed utterances by their speech acts in order that the program might recognise the speaker's plans and so provide the most helpful response to an utterance by recognition of the speaker's intent. Allen gathered a corpus of exchanges between members of the public and ticket collectors and information clerks at the Toronto railway station. He noticed that often the railway official would provide more information than was strictly asked for:

> Patron: When does the Montreal train leave?
> Clerk: 3.15 gate 10. (Allen 1983: 107)

He made the following three assumptions about conversations, in accordance with this observation:

> (1) People are rational agents who are capable of forming and executing plans to achieve their goals.
>
> (2) They are often capable of inferring the plans of other agents from observing that agent perform some action.
>
> (3) They are capable of detecting obstacles in another agent's plans. (op. cit.: 107-8)

---

[1] For a discussion of these two approaches and an attempt to reconcile them under one theoretical architecture of information states for task-oriented instructional dialogue, see TRINDI 2001. I shall discuss TRINDI briefly in Section 6.2.4.

So the clerk responded with the unasked for information about the location of the departure because he was able to infer the patron's plan – i.e. that the patron wished to catch the train, and therefore would need to know where the train would leave.

Allen sums up his plan based approach in the following way:

> The world is modelled as a set of propositions that represent what is known about its static characteristics. This world is changed by actions, which can be viewed as parameterised procedures. Actions are described by preconditions, conditions that must hold before the action can execute, and by effects, the changes that the action will make to the world. Given an initial world state W and a goal state G, a plan is a sequence of actions that transforms W into G. (op. cit.: 111)

This idea is largely borrowed from work by Fikes and Nilsson (1971) on a problem solver, STRIPS (**ST**anford **R**esearch **I**nstitute **P**roblem **S**olver), to solve robotic problems. They wanted to be able to give a robot a general command such as "Push three boxes together" and have the robot break the overall problem into achievable sub-goals. It did this by trying to find a contradiction to the overall goal, and using the incomplete proof to provide sub-goals. This problem solver worked quite well in a limited domain, with a small number of actions, defined as well-formed-formula schemata.

Allen (1983) adapts this idea to linguistic actions in order to be able to recognise the sub-goals necessary for the speaker's main goal to be achieved. The action is represented in the form of a schema, consisting of a name, a set of parameters and sets of formulas (possibly null). These formulas are divided into the following classes:

- *Preconditions*: Conditions that should be true if the action's execution is to succeed.

- *Effects*: Conditions that should become true after the successful execution of the action.

- *Body*: A specification of the action at a more detailed level. This may specify a sequence of actions to be performed, or may be a set of new goals that must be achieved. (op. cit.: 118)

A speech act is defined as an action that has as its parameters a speaker, a hearer, and a propositional content. Its preconditions and effects are defined in terms of the speaker's beliefs and wants. So the action of informing is given as the following schema:

INFORM (speaker, hearer, prop)[2]

precondition:	KNOW (speaker, prop)
effect:	KNOW (hearer, prop)
body:	MUTUALLY BELIEVE (hearer, speaker, WANT (speaker, KNOW (hearer, prop))).

Allen's program creates plans by finding the set of all partial plans that map the initial state onto a goal state. He describes and specifies two possible goal states in the train domain: that the speaker wishes to BOARD a train, or the speaker wishes to MEET a train. By identifying the speech act (INFORM, REQUEST, etc.) a partial plan is constructed bottom-up by using certain plan inference rules from the observed action (speech act) – this Allen calls the *alternative*. A second partial plan is constructed top-down using certain plan construction rules from an expected goal (either BOARD or MEET) – this is called the *expectation*.

So with the utterance

> 'When does the train to Windsor leave?'

the program parses it into an appropriate semantic representation and assigns possible speech act types to it using a variety of criteria such as sentence mood (syntactic structure of sentence). Having recognised this as a request (to inform of a reference), the system tries to construct a suitable plan to identify the intention. The alternatives are expanded using inference rules, and expectations are expanded using planning rules. By using forward and backward chaining methods, using rating heuristics to assess the probability of any partial plan's being correct, the search continues until the expectation and alternative plans 'meet', thus signalling that a plausible plan has been identified.

Once the speaker's plan has been identified, obstacles to the plan's execution are detected and the reply will take these obstacles into consideration, attempting to remove them – hence the generation of more information than was strictly asked for.

Allen's plan-based approach to speech act recognition is appealing for a number of reasons. It attempts to integrate speech acts within a Gricean pragmatic framework, and more importantly makes no distinction between direct and indirect speech acts, treating them uniformly when formulating plans, and allowing utterances to perform more than one speech act simultaneously.

---

[2] Note that this bears a striking resemblance to some of the felicity conditions defined by Searle for the speech act ASSERT.

However, Allen's model and especially his heuristics are domain specific; it is more than unclear whether this approach could be extended to an unlimited domain without the combinatoric complexity involved making the model computationally intractable. It is difficult, therefore, to see how this model might be expanded to cope with general conversation. The domain (information desk at a railway station) Allen uses to construct his model is a highly structured one in which there is generally a question-answer expectation. This approach might work if one were thinking of setting up a simple question-answer database to replace the clerk at a train station or some similar situation. But it is doubtful that the subtleties of everyday conversation could be represented in this way.

6.1.2 Planning Utterances

Another related computational model has been built by Appelt (1985) to generate appropriate speech acts. His system KAMP (**K**nowledge **A**nd **M**odalities **P**lanner) has as its domain the assembly and repair of complex electro-mechanical devices. Whereas Allen's program concentrated on inferring speaker plans from an utterance, Appelt is concerned with how a person's beliefs and intentions cause a speaker to produce a particular utterance.

KAMP is a hierarchical planning system that uses a non-linear representation of plans known as a procedural network. There are four levels of linguistic actions that make up the hierarchy:

Illocutionary acts (such as INFORM and REQUEST).
Surface speech acts (such as ASSERT, COMMAND and ASK).
Concept activation (such as DESCRIBE and POINT).
Utterance acts.

The system is given a top-level goal, and the linguistic levels are expanded downward by KAMP, beginning with the illocutionary act and ending up with the utterance act.

Appelt runs through an example, where the computer, called Rob (short for robot) has to solve the problem 'Detach the pump from the platform'. So the top-level goal will be ¬*Attached (PU, PL)*. KAMP tries to show that the goal is already satisfied in the world W_0. Since it is not, KAMP goes on to try and find some action that will have as its effect that the goal will be achieved. It finds that REMOVE will have the desired effect, but that the action of removing requires a human agent. As Rob is a non-human agent, KAMP has to plan to use John (the apprentice) to REMOVE the pump from the platform. Thus KAMP looks to see what the conditions are for John to perform such an action. It comes up with two further sub-goals: that John must intend to REMOVE the pump and that he must be in the same location as the pump. As the second condition is fulfilled, attention is switched to the first. KAMP then attempts to show that John wants to REMOVE the pump in the world W_0. As this is not so, it searches for

the conditions necessary for John to want to REMOVE the pump and finds that the action REQUEST will have the desired effect. So KAMP plans to REQUEST that John will REMOVE the pump. The plan has been completed at the highest level of abstraction (i.e. the illocutionary act).

KAMP goes on to plan the surface speech act of COMMAND. There is no further linguistic planning at this stage, but before unifying this with the grammar, the full conditions of REMOVE have to be expanded. So KAMP searches for the conditions of fulfilment for the act REMOVE, and finds that in order to remove an object, the bolts holding it need to be unfastened, and in order to unfasten them a tool is needed. After examining what Rob knows about John, KAMP has to plan to tell John what tool it is that he needs to remove the pump (a wrench) and where that tool can be found (in the tool box).

Once the cycle of expansion and criticism (Sacerdoti 1977) is completed, and KAMP has worked out exactly what John needs to know in order to perform the action REMOVE, the functional description associated with each of the surface speech act 'nodes' (i.e. which tool and location of tool) is unified with the teleological grammar[3]. The following utterance act is then performed:

'Remove the pump[4] with the wrench in the tool box.'

KAMP seems to perform fairly well and Appelt suggests that from the results of testing the program, this approach constitutes a feasible approach to utterance production.

The work of Allen and Appelt is representative of the overall approach to speech act theory by researchers in artificial intelligence. In many ways, their implementations have highlighted the problems faced by those dealing with speech act theory in general. Allen's work is exciting as it is indicative of the representational possibilities of a purely pragmatic approach. Appelt's work shows clearly how utterances can be planned by reference to their speech act.

However, both these models fail to take some vital concepts into account (for the purposes of this work). Speech acts are never just randomly used. They fit into discourse structure in well-defined ways, and are regulative of it. It is not enough to say that there are certain speech act types that are used, and then go no further. Speech acts have the effect of constraining the range

---

[3] This grammar, which Appelt calls TELEGRAM, is a schematic grammatical representation of the surface speech acts; filling in the schema produces the final utterance. See Appelt 1985: 135-49 for details of this example and a full description of the grammar.

[4] The extra information that the pump is attached to the platform is unnecessary and is therefore omitted from the request. John already knows that the pump can only be attached to one thing and that it is attached to the platform.

of further speech acts that will follow in the conversation. To completely ignore these relations between speech act utterances is to create many problems when examining speech act theory in the context of conversation, where they become evidently important.

6.1.3 Domain and Discourse Plan Recognition

Following on from earlier work in plan recognition (as discussed above), Litman and Allen (1984, 1987, 1990) distinguish between domain plans and discourse plans, which are themselves domain-independent. The former specify how tasks should be carried out, whereas the latter are generated to control the flow of communication and are generated in response to the execution or discussion of the domain plans. Litman and Allen justify the separation of these two planning behaviours for these reasons:

(1) They are different processes; whereas the domain plans refer to the action being performed, the discourse plans refer to the domain plans (i.e. how by uttering a string of words I can contribute and bring myself closer to the accomplishment of a domain plan).

(2) The elements in a discourse plan need not refer to any of the elements of the domain plan that was responsible for its generation.

(3) The planning of discourse elements does not necessarily follow the execution structure or the sequential ordering of elements in the associated domain plan.

Litman and Allen develop a library of domain plans, and a smaller group of discourse plans to help carry out the domain plans; they do not claim to have provided a comprehensive set of discourse plans, but have only included those that are necessary for their purposes. They argue that these are adequate enough to cover a large variety of different discourse situations.

Litman and Allen classify discourse plans into the following three categories:

(1) *The Continue Class*: Plans that allow the effective continuation of a plan in execution. An example of the continue class of plan is TRACK-PLAN, which is triggered by a non-linguistic executing task within the domain plan, so that discourse is generated to explain one of the different steps in the task. Uttering something like, ‘Here’s five pounds’ when paying for a ticket at a train station is an instance of a plan for continuing.

(2) *The Clarification Class*: Plans that involve the clarification or correction of a plan in execution. An example of a clarification plan is IDENTIFY-PARAMETER, which is the discourse plan that is generated when there is a lack of information that prevents the underlying domain plan from being executed. If I need to catch a train, but do not know which platform it is leaving from, I will generate an IDENTIFY-PARAMETER plan to ask the location of the platform. An example of a correction plan is CORRECT-PLAN, which

specifies the correction of a plan when unforeseen events happen while the plan is executing. For example, if I try to pay for a train fare with a five-pound note, but the fare costs more than this amount, the ticket seller will have to formulate a CORRECT-PLAN in order to change my incorrect assumption of the ticket price.

(3) *Topic Shift Class*: Plans that introduce new, or modify existing, plans. An example of a topic shift plan is INTRODUCE-PLAN, which presents a new plan into the discourse context that has not been part of any prior discussion. For example the utterance, 'I want to buy a return ticket to Liverpool please' introduces to the ticket seller my plan to buy a ticket to Liverpool (presumably so that I can get on a train and actually go there).

These discourse plans are encapsulated as STRIPS plan operators. So, for example, the schema for INTRODUCE-PLAN is represented as (from 1990: 375):

INTRODUCE-PLAN (speaker, hearer, action, plan)
decomposition: INFORM (speaker, hearer, WANT (speaker, action))
effect: BELIEVE (hearer, WANT (speaker, plan))
constraint: STEP (action, plan)

As well as the libraries of discourse and domain plans, Litman and Allen also include a library of speech act schemata based on the work of Cohen and Perrault (1979), Cohen et al. (1982) and Allen (1983, 1987). With these as a basis, they extend the earlier models of discourse to be able to handle confirmations, corrections, and interruptions in a restricted domain. The novelty of Litman and Allen's work mainly consists in the separation of discourse and domain plans, and the dependence of the former on the latter. They introduce a simple plan recognition system, and propose that the consideration of certain linguistic cues might further aid in the recognition of the different classes of discourse plans being performed.

Other researchers would make the role of linguistic cues more central to the structure of discourse. Grosz and Sidner (1986), whose work I shall consider in Section 6.2.1, based a key part of their theory of discourse on just such an assumption.

6.1.4 Linguistic Information and Plan Recognition

Hinkelman (1989) and Hinkelman and Allen (1989) produce a model of speech act recognition that attempts to marry a syntactic and semantic, language specific linguistic analysis with a plan-based approach to intention. Hinkelman also develops a simple theory for the identification of plan-based conversational implicatures, which uses knowledge of interpretation in order to infer any extra information that may be derivable from the recognition of the intention in utterance production.

Hinkelman uses a surface linguistic analysis to account for language specific interpretations of utterances, especially those that give rise to indirect speech acts. She gives the following two sentences as examples of the kind of lexicon level clues to speech act interpretation (1989: 12):

(1) 'Can you speak Spanish?'
(2) 'Can you speak Spanish please?'

In the first example, the utterance can be analysed as either a request (for the hearer to speak Spanish) or a yes-no question (when the speaker just wants information and therefore expects either a positive or negative answer). In the second example however, the addition of the word 'please' indicates that the force of the utterance should be understood as a request rather than a simple yes-no question. This effect is not only noticeable with the use of the word 'please', e.g.:

(3) 'Are you able to speak Spanish?'
(4) 'I hereby ask you to tell me if you can speak Spanish.'

So, while the paraphrase of (1) in (2) is clearly a request, the paraphrases in (3) and (4) are clearly questions about the hearer's ability; although arguably (3) might be answered not just by 'Yes' or 'No', but also by 'Por supuesto que sí' ('Yes, of course') by the action of speaking Spanish, as well as by providing an affirmative answer (in Spanish).

However, linguistic features of an utterance are patently not sufficient indicators of speech act interpretation in themselves (as I have discussed in detail in the earlier sections of this dissertation). Look at the classic example 'It's cold in here' again, which can be interpreted as a statement, request or interrogative in the appropriate context without any need to change the surface linguistic form. Hinkelman suggests that a theory of speech acts requires a combined approach in order to cover the identification of a wide range of speech acts. Here again the semantics/pragmatics theoretical divide can be discerned, but in the setting of plan-based computational approaches to speech act recognition. Hinkelman and Allen (1989) note that plan-based approaches that cannot handle both aspects, subscribe (whether consciously so or not) to the hypothesis that there is one and only one 'correct' speech act being performed by an utterance in discourse, and therefore only one possible 'correct' interpretation.

Previous plan-based approaches to the recognition of indirect speech acts would first try to interpret the utterance literally, and only after encountering some inconsistency with this interpretation would reasoning be used to infer the act that was actually intended indirectly. Hinkelman and Allen argue that the surface (syntactic) structure of an utterance is often a very

poor indicator of the utterance intention[5]. Other factors (often contextual) govern the identification of speech acts. One argument put forward against inferring indirect speech acts in this way is that, if one could generalise the identification of indirect forms of utterance, then they would be used in the same way in different languages. Searle (1975) notes that the literal translation of a sentence such as (1) cannot be used as an indirect request in Czech. So one can deduce from this that directness (and literality) is not culturally invariant. In Spanish (or at least, Argentinian Spanish) for example, indirect forms are much less likely to be used, as indeed are the phrases 'please' and 'thank you'. They have more tolerance for what in British English would be termed 'impolite' forms of address, and if anything perceive the English as over-courteous – almost too polite to be sincere.

With this language specificity in mind, Hinkelman and Allen suggest that the interpretation of some utterances is inherently ambiguous even when constrained by their linguistic structure. The ambiguity is resolved by contextual, language and cultural dependent information. Thus an utterance is parsed and the syntactic structure (presumably 'mood' or sentence type would be of primary interest here) and the semantic content (cue words and phrases) are used to cut down the number of different potential speech act interpretations. If there is still any ambiguity, it is then resolved by resorting to plan-based inference and reasoning techniques.

Hinkelman and Allen capture the use of conventional speech act indicators on the surface form of an utterance by matching the results of linguistic parsing with rules for the production of such conventional acts. Each rule corresponds to some syntactic or semantic features of such utterances, which then 'suggest' an interpretation. For example, the rule for identifying a simple yes-no question is (op. cit.: 8):

IF *Mood* is Yes-No-Question THEN
    Possible interpretations: any speech act
    Suggested interpretations: a yes-no question

The rule for converting the interpretation of a yes-no question above into a request, when phrased in the form 'Can you *X*?', is (op. cit.: 9):

---

[5] This is backed up by empirical studies such as those carried out by Hassell and Christensen (1996). They claim that, from the results of their research into the use of indirect utterances in three means of business communication – face-to-face, telephonic, and email – people are at least twice as likely to formulate non-assertive expressions indirectly than they are to do so directly.

IF *Mood* is interrogative AND *Subject* is "you" AND *Voice* is active
AND *Aux* is "can" THEN

Possible interpretations:	any speech act
Suggested interpretations:	a request to do the act corresponding to the verb phrase (*X*)

These production rules can be applied incrementally in order to restrict increasingly the range of possible acts under consideration. If there is only one option left, all well and good – the speech act has been recognised. Otherwise the set of potential speech act interpretations is passed to a plan recognition module for further processing.

The plan recogniser is used in two cases: to select the most likely candidate from the set of speech acts suggested by the linguistic parser module, and to infer an interpretation when none of the linguistic rules are applicable to the utterance. The model uses traditional plan recognition techniques to come to a final interpretation.

The same process is applied to infer implicatures when some of the knowledge in the identification of a speech act is missing. While speech acts whose preconditions are false are dismissed out of hand, it may sometimes occur that the information needed to make such a decision is incomplete or unknown. In this instance, the candidate speech act is not rejected immediately, and it may even be chosen as the correct interpretation in the current context in the absence of a more appropriate candidate. If it is selected, then the unknown part of the precondition is 'implied' (from the success of the speech act) and added to the background beliefs of the system. This seems like a good idea in principle as long as it is possible to backtrack at a later point in the discourse (when for instance it becomes obvious in some way that the implicature was incorrect). Implicatures, as defined by Grice must be defeasible (as must be the interpretation of speech acts also).

In short, Hinkelman and Allen outline a model that draws together a plan-based model and a heuristic-driven procedure for determining speech acts from linguistic features of utterances; this has the advantage of increasing efficiency because of the added constraints upon the search space for appropriate speech act identification, as well as providing an integrated account of conventional indirect forms of speech. They include as well, a simple method of aggregating implicatures to the background knowledge of participants.

6.1.5 Conversation Acts

Traum and Hinkelman (1992) extend traditional approaches to intention-based speech act recognition by dealing with structures at lower levels of abstraction, such as grounding, turn-taking and argumentation architectures. Their work on conversation acts is implemented within the TRAINS dialogue system (Allen et al. 1994, which I shall discuss briefly in Section 6.1.6).

They argue that, contrary to the assumptions commonly made in other approaches, the following amendments need to be made:

(1) It is not the case that speaker and hearer automatically mutually assume that the hearer has heard and understood what the speaker says. A model of conversation must therefore include an account of error recovery and feedback/backchannel production. In other words, that there must be some aspect of dialogue control embedded and built into the system.

(2) Speech acts are not single-agent plans – or at least, the useful accomplishment of a speech act in dialogue usually requires some form of uptake by the hearer. Therefore it is argued that speech acts are a multi-agent phenomenon (or as Clark 1996 puts it, a joint activity – I shall be returning to this idea in Chapter 8).

(3) Utterances do not necessarily perform a single speech act within the discourse. Not only this, but an utterance can have an effect on more than one level as well: dialogue control or turn-taking signalling for instance.

I have already established these points in earlier discussions in this dissertation, however it is true that such considerations have largely been missing from plan-based models of speech act theory. Traum and Hinkelman introduce a new framework of what they call 'conversation acts', which claims to deal with the actions performed in dialogue in a more general manner than other accounts. They describe four different levels of action in dialogue:

(1) *Turn-taking acts*: These are at the lowest level and allow participants to KEEP, RELEASE, TAKE, or PASS UP their turns. Thus control of the dialogue is passed from speaker to speaker.

(2) *Grounding acts*: These are acts that assure participants that some common ground has been established between then. Types of grounding act are: INITIATE, CONTINUE, ACKNOWLEDGE, REPAIR, REQUEST-REPAIR, REQUEST-ACKNOWLEDGE and CANCEL.

(3) *Core speech acts*: These correspond to the more traditional accounts of speech act theory. The core acts are at a higher level than turn-taking and grounding. Examples include: INFORM, ACCEPT, REQUEST, YN-QUESTION, WH-QUESTION, SUGGEST, EVAL, REQUEST-PERMISSION, OFFER and PROMISE.

(4) *Argumentation acts*: These come above the core acts and cover combinations of speech acts, which unite together to achieve larger goals in the conversation. So for example, a speaker might group together INFORMs and QUESTIONs in order to try to CONVINCE some other participants to perform some plan on the speaker's behalf. Argumentation acts constitute the overarching plan in the production of the current sequence of utterances (equivalent

perhaps to Grosz and Sidner's 1986 discourse segment purpose, which I shall be covering in Section 6.2.1).

The four types of act described above form a hierarchy of intention, with turn-taking and grounding acts contained in speech acts, which are in turn contained by argumentation acts. There are different methods for recognising the performance of these different types of act. Turn-taking is largely dependent on the structure of the social setting or occasion of the conversation. There is a scale of rigidity with turn-taking procedures whose application depend on the formality of the conversational situation. In an interview one would expect a strict respect for turn observation, whereas in a party one might feel able to usurp a turn and interrupt at an appropriate point even if the speaker has not finished what he is saying. In Traum and Hinkelman's system, turns are taken up by the indicated participant at the next neutral transition point in a discourse. Grounding acts are recognised by the use of a finite state transition network, the traversal of which gives all the 'legal' sequences of grounding act allowable in a conversation. Core acts are identified by syntactic analysis and plan recognition techniques (Hinkelman 1989). Argumentation acts are also recognised by the application of dialogue act schemata, and the matching of cue words and phrases (e.g. 'So', 'Therefore', 'Now then', 'Yes but', etc.) to their respective conversational functions.

6.1.6 The TRAINS Project

The TRAINS system (Allen et al. 1994) incorporates all of the improvements to the basic plan-based design of speech act representation that I have discussed so far in this first section of the chapter. The TRAINS project, which was first started in 1990, was an attempt to develop a natural language interface to allow a problem-solving computer program to assist and co-operate in the planning of the transportation of goods by train between factories and warehouses. The system is designed to be a planning agent, which plans co-operatively and in real time with the user to schedule the movement of materials. It was developed as a 'toy system' on an overly simple task to show the potential of automated dialogue systems, and was not intended for real use.

The TRAINS system was first developed by collecting a corpus of simulated planning interactions, with a human participant taking the part of the system at first. The user is given a number of goals to achieve and utilises the 'system' to help to plan solutions. For example, a typical goal might be: 'Get one boxcar of oranges to Bath by 8am'.

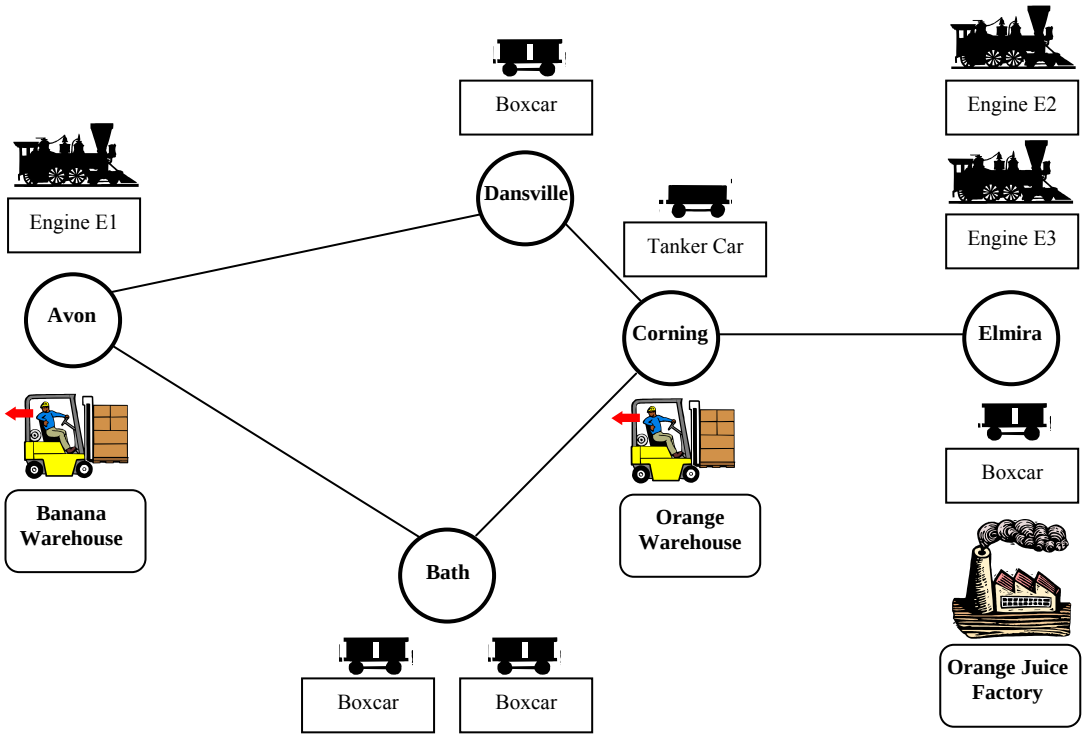


Figure 6.1 A simple TRAINS world map.

The participants have access to a map (as shown in Figure 6.1 for example), which displays the current positions of all the vehicles and engines (as well as all the factories and warehouses). There are two possible solutions for the first sub-goal of getting an engine and a boxcar to Corning in order to pick up the oranges: (1) engine E1 could pick up a boxcar from Dansville or Bath, or (2) one of the engines, E2 or E3, at Elmira could hitch up with the boxcar there. The planning agents should choose (2) as the more time efficient option. From Corning, planning the final leg to Bath is trivial. From the collection of this corpus of planning interactions, rules about how such collaborative behaviour occurs were formalised and implemented in a computer program. I do not want to go into too much detail about how this was done, as it is not really relevant to the discussion, but the following speech acts are identified and executed as types of event within the TRAINS-93 system:

T-INFORM	The speaker aims to establish a shared belief in the proposition asserted.
T-YNQ	The speaker asks a yes-no question, creating an obligation for the hearer to respond.
T-CHECK	The speaker is verifying that a certain proposition is true (that the speaker already suspects is true).
T-SUGGEST	The speaker proposes a new item (action, proposition) as part of the plan.
T-REQUEST	The speaker aims to get the hearer to perform some action. In the TRAINS domain, this is treated like a suggest, with the addition of an obligation on the hearer to respond.

T-ACCEPT	The speaker agrees to a prior proposal by the hearer.
T-REJECT	The speaker rejects a prior proposal by the hearer.
T-SUPP-INF	The speaker provides additional information that augments, or helps the hearer interpret some other accompanying speech act.

Planning in the TRAINS domain is extremely complex as it involves taking into account synchronous, variable duration and situation specific actions. In order to keep the system manageable therefore, the belief model had to be kept as simple as possible. The facts in the domain are assumed to be directly accessible to all participants, which in essence equates to mutual knowledge. However, plans are not, at first, expected to be common to both system and user; it is only after a proposed plan is accepted that it becomes a part of the domain knowledge (see Figure 6.2).

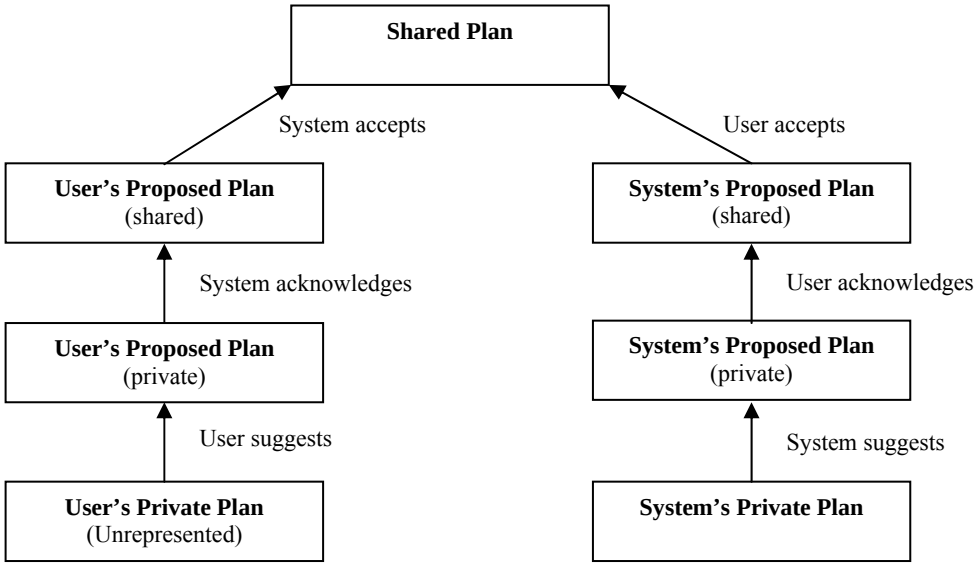


Figure 6.2 The different plan modalities for handling suggestions in TRAINS.

One of the greatest strengths of the TRAINS project is that it demonstrates a working system, an implementation of a plan-based approach that works in real time (albeit in a very restricted domain). However, arguably it is simplified to such an extent that it cannot be seen as a realistic model of human interaction. Conversations in TRAINS are always represented as co-operative, and the contents are almost completely devoid of knowledge of the real world. The large quantities of data and processes that need to be modelled in order to infer things in the real world are missing (for reasons of computational tractability). Most of the discourse centres on the restricted domain of a map of the locations of factories, and the routes of the railways (see Figure 6.1).

Referring expressions are usually related to features of this map, so that there is little ambiguity respecting their interpretation. Not only is the dialogue always co-operative, but also the belief

spaces of the user and system are extremely, and predictably, similar; so there is no need to include dialogue control aspects of language in the TRAINS model (although in later versions of the TRAINS system this has been added).

The claim of researchers working on the TRAINS dialogue system is that although the application is specific to the task of transporting goods by train, the system itself is purported to be a platform for demonstrating the potential of *general* language processing and plan reasoning techniques by its implementation in a specific domain. The modules that have been developed are allegedly therefore generic in nature and extensible to other domains.

This is rather difficult to believe simply on the basis of how the task itself was chosen. Allen et al. (1994) describe how other tasks that were considered in the implementation of this project to build a plan-based natural language interface were rejected as either too complicated (as in the case of a librarian's assistant) or not covering a rich enough range of dialogue behaviours (as in the case of a robotics-controlled model railway layout). They justify their choice of application as being dictated by the constraints of finishing the project within a specified timescale. While I do not wish to detract from the achievements of this approach, I would be very wary of admitting claims of generality, even at the task and domain-restricted level. It seems like no simple task to use this system for any other task-oriented purpose.

The problem is, humans do not confine themselves to a fixed manner of communication even in highly constrained frameworks. A system that cannot adapt to general rules of conversation will always be subject to weaknesses. Computationally this leads to a (currently unsolvable) dilemma – the more general the model, the less manageable.

Allen et al. (1994) themselves recognise a number of problems with the earlier TRAINS system as described above:

(1) The system cannot handle speech input. Negotiation is carried out via a keyboard and screen.

(2) There is no intelligent recovery when the system is not able to process and understand the input. The parsing process is an all-or-nothing event; there is no attempt to extract partial meaning from an utterance, and no repair mechanism.

(3) The TRAINS system relies on the user's initiative alone – a collaborative planning assistant might need to communicate a problem concerning a plan about which the user is unaware (i.e. if there are leaves on the line between Dansville and Corning causing a delay for trains along that route, a plan that directs a train that way may fail in its object of reaching a destination before a certain time).

The most recent version of TRAINS, TRAINS-95, as described in Ferguson et al. (1996), addresses some of these issues identified in the earlier system (the numbering of the solutions matches the numbering of the problems above):

(1) TRAINS-95 uses a more complex map (shown in Figure 6.3, copied from Ferguson et al. 1996), than the earlier system, although at the same time, the problem has been considerably simplified. It is no longer about the transportation of goods, but about the movement and scheduling of trains to different locations. Input to the program is made either by speech, keyboard, or by clicking with a mouse directly on the graphical display (i.e. input is multi-modal). All these actions are viewed and analysed as 'linguistic' forms of communication, and are treated in the same way by the system. This reflects the kind of behaviour that might take place between humans; often when explaining or planning a route it is natural to point out locations and course of the railway on a map.

Figure 6.3 TRAINS-95 system map display.

So, the system is able to deal with multi-modal interaction. Little is said by Ferguson et al. (1996) however about whether the user can utilise more than one of these modalities at the same time, nor how the system might combine input modalities to come up with one overall 'message' from the user.

(2) TRAINS-95 uses a parser that is now more robust, so that even partial utterances are dealt with and likely speech acts assigned. This is partly achieved by the implementation of a dialogue context, so that the goals specified at the beginning of the current exchange are used for grounding the subsequent utterance interpretations (e.g. knowing we are talking about an engine allows the interpretation of the indefinite reference in 'Send *it* to Chicago'). There is however, still no method described for error recovery if the system assigns an incorrect interpretation to an utterance.

(3) The TRAINS-95 system allows mixed-initiative[6] planning to take place. If there is information available to the system that might pose an obstacle for the successful completion of a journey plan, then the system takes the initiative to inform the user of the potential threat to the goal, so that an alternative route can be planned if necessary. It is interesting to note that traditional approaches to plan generation do not really work well in mixed-initiative dialogues. This is because the straightforward solution to the original goal as stated by the user, is rarely adhered to exactly without modification due to the new information that the system provides during the course of the interaction.

In order to encourage collaborative interaction between the system and user, the domain reasoner in TRAINS-95 has been deliberatively weakened so that the user has to interact in order to overcome the system's shortcomings. The route planner only plans routes that are at most four steps long, and the path chosen is random. It is only the system that knows of delays (adverse weather conditions, a bottleneck at a train terminal, etc.), so that the system also is forced to communicate these difficulties to the user. Of course, a more efficient route planner could have been implemented, but then there would have been no motivation for mixed-initiative collaboration. Ferguson et al. argue that:

> …not only is it unlikely that we will ever be able to build such a reasoner for a realistic domain… we claim that such a system is not necessarily appropriate for mixed initiative planning. (1996: 73)

This seems like a backward step. It is unclear to me how weakening a computational model in this way simply to allow more 'interesting' interactions to occur is of any practical use. If the aim is to produce a tool to help a human planner, then surely the system planner should come up with the best plans possible, and be able to explain its choices to the human user. To artificially create a situation for collaboration between human and computer, apart from mere curiosity at the results, might almost be said to be a

---

[6] I mention mixed-initiative planning here, but do not intend to discuss it at length. For detailed discussion, see Boye et al. 1999, Chu-Carroll and Brown 1997, Churcher 1997, Lee 1997, Smith and Gordon 1997, and Williams 1996.

> waste of time. The most damning admission made by Ferguson et al. here is that such a system *cannot be scaled up* to a more realistic (and therefore by implication, more complex) application. I would suggest that part of the difficulty with the TRAINS system as a whole, is that it is perhaps trying to accomplish and satisfy too many goals at once. The task that has been chosen is in itself one of the most complicated to model computationally; scheduling (of transport, of computer jobs, etc.) and constraints satisfaction are large and rich areas of research in themselves.

Although TRAINS-95 is in many ways a considerable improvement on the earlier system, the planning domain is extremely limited. The problems to be resolved by the user and system are mainly just those of transferring a train from one location to another. It is difficult to see how the system could possibly be implemented so as to be of any practical use in the real world.

6.1.7 Conclusions about Plan- and Intention-Based Analyses

Intention-based approaches to discourse modelling and speech act recognition take as their fundamental assumption that the coherence of a discourse can be represented from the point of view of the intentions (as well as the plans, goals and beliefs) of the participants. As such, the focus is mainly on the recognition and modelling of the plans, goals and beliefs of participants. There are a number of drawbacks to this approach as a whole, not the least of which is the enormous computational effort required to process the information to make the necessary deductions and inferences from the data, even in 'toy systems' such as TRAINS. The processing effort increases in a non-linear fashion with the addition of every new fact or rule, because of the increase in size of the search space of possible combinations. Intention-based models have been restricted to task-oriented dialogues with good reason: the more constrained the field of interaction, the more likely the model is to work well and produce good results.

Modelling the beliefs of a participant in a discourse is also difficult from the viewpoint of nested beliefs, which if represented can lead to an infinite regression of beliefs. Once again this has been combated by the introduction of (arbitrary) limits to the depth of nesting, and by the ascription of beliefs at need, rather than by exhaustive generation.

Apart from intention-based models of speech acts, there is a second type of approach, which is based on the linguistic theories of dialogue structure. It is to the application of these by various computational linguists to which I now wish to turn.

6.2 Structure- and Cue-Based Analyses

As we have seen, intention-based approaches have had limited success in defining the identification of speech acts. Structural methodology looks at interaction from a syntactic rather

than a semantic point of view. Amongst the earliest proponents of this approach are Grosz and Sidner, who suggest a theory based on discourse structure.

6.2.1 Discourse Structure Theory

Grosz and Sidner's (1986, 1990) proposed theory of discourse structure is an attempt to unify two different strands of research in the computational modelling of discourse: (1) focusing and structure in discourse (Grosz 1977, 1978a, 1978b, 1981, Reichman-Adar 1984 and Reichman 1985), and (2) intention recognition in discourse (Cohen et al. 1982, Allen 1983, Sidner 1983, 1985, Litman 1985, Pollack 1986, Carberry 1990 and Cohen and Levesque 1990). It is the structure-based theories that are most prominently represented in their research.

Grosz and Sidner propose that discourse is made up of three subcomponent structures: linguistic, intentional and attentional.

Linguistic structure: The first component of discourse structure theory is the structure and sequence of utterances in a discourse. This is not so much how utterances are related to each other, but how utterances combine together to form larger discourse segments. The utterances are relevant to the discourse segment to which they belong, and the discourse segment is relevant to the discourse as a whole. So, discourse segments are units in a discourse, just as utterances are units in a discourse segment, and words are components of an utterance. Utterances that follow each other in a discourse may fit into the same discourse segment, or belong to different discourse segments. Similarly, utterances that are separate may correspond to the same segment. The discourse segments themselves may be embedded within each other, so forming larger discourse units.

Grosz and Sidner suggest that there is a two-way relationship between the utterances and the discourse segments. Utterances are used to signal changes in the higher level structure of the discourse segment by the use of cue phrases (what Reichman 1985 calls clue words) such as 'actually', 'in the first place', 'remember', 'moving on', 'anyway', etc. and also by the use of intonation. However, the interpretation of the utterances themselves may also crucially depend upon the type of discourse segment within which they are performed. In other words, the same utterance, it is argued, may be performing different roles in different discourse segments.

The linguistic structure of discourse proposed by Grosz and Sidner appears to be a reformulation of earlier work on the rhetorical structure of textual coherence in written texts carried out by Mann and Thompson (1983). Grosz and Sidner are not the only proponents of the organisation of discourse units in such a way; Reichman-Adar (1984) also posits a similar theory. However, there is little agreement as to how the segmentation of a discourse should be achieved, nor even where the boundaries of a segment may lie (although it is suggested by

McKevitt et al. 1992, that the boundaries of discourse segments may be indicated by the use of linguistic expressions as listed above). Grosz and Sidner provide no method to delimit or define the recognition of a discourse unit; there is no indication either of the number of different types of unit there may be.

There are psychological experiments (Mann et al. 1975) whose results show that people do indeed tend to break up discourse into units in similar ways. Although Grosz and Sidner define discourse in terms of segments and coherence ties (as identified in linguistic markers or cue words and phrases) that relate one segment to another, they fail to provide a general taxonomy of these segments and how they can be identified.

Intentional structure: The second tier in the discourse structure theory is the structure of the various intentions or 'purposes' of the participants. Knowledge of the purpose of a discourse will help to distinguish between coherent and incoherent utterances. Although a participant may have several underlying intentions for engaging in discourse, Grosz and Sidner specify that all discourses have a foundational discourse purpose (DP), which motivates the performance of the interaction as a whole, and underlies the transmission of the information being communicated. The DP is intended to be recognised by other participants in the discourse, and as such is closely related to Grice's idea of communicative intention. The DP itself is composed of sub-units of intention, so that for every discourse segment there is a discourse segment purpose (DSP) that led to its initiation. The DSP relates the discourse segment to the overall DP and contributes to its satisfaction. We can see that the intentional structure, intuitively, mimics the linguistic structure as described above, with DSPs embedded within the DP. Grosz and Sidner (1986: 179) give the following list of types of intentions that serve as DPs/DSPs:

- Intend that some agent intend to perform some physical task.
- Intend that some agent believe some act.
- Intend that some agent believe that one fact supports another.
- Intend that some agent identify an object.
- Intend that some agent know some property of some object.

Grosz and Sidner identify two types of relationship that may hold between DSPs: *dominance* and *satisfaction-precedence*. One DSP is said to dominate another if in the course of satisfying the other, the original discourse purpose is furthered. Thus, if DSP2 *contributes* to DSP1, then DSP1 *dominates* DSP2. This creates a *dominance hierarchy* or chain, which partially orders DSPs within the overall DP. This ordering may be crucial to the successful achievement of the DP, in which case DSP2 is said to *satisfaction-precede* DSP1, because DSP2 must be satisfied before DSP1 can be. Satisfaction precedence could be of significance to the order in which DSPs are realised and satisfied, especially in task-oriented dialogues.

Grosz and Sidner propose that studying the interrelationships between the various discourse intentions in this way is important because, as Wittgenstein (1953: paragraph 23) also argued, there cannot be a finite set of such purposes; therefore the recognition of DPs cannot be based upon the definition of a fixed number of intention frameworks (as is often suggested by other research). This is all very well, but the definition of these relationships does not really further the cause of recognising the intention of a speaker in producing a particular utterance (and thereby understanding his meaning).

Attentional structure: The third structural component of Grosz and Sidner's theory is the focus of attention within the discourse. The attentional state is a property of the discourse itself, and not of the participants. This is important to note, because while the former is feasibly modelled, the latter is not. The attentional state serves to maintain a dynamic record of those entities in the current discourse that are most salient at any point in time. The following abstract structures are assumed to exist (as given in McKevitt et al. 1992: 345-6):

(1) Focus space: a set of these models attention.

(2) Transition rules: model changes in attention state. These specify conditions for adding/deleting focus spaces.

(3) Focusing structure: the collection of focus spaces available at any one time that are salient at each point.

(4) Focusing: the process of manipulating focus space.

Focus spaces are organised in a data structure known in computational terms as a *stack*. A new focus space is pushed on top of the stack, and represents the current focus of the discourse. When the DSPs associated with the current focus space is (eventually) completed, the focus space is popped off the top of the stack, and the previous one becomes the most relevant focus space (unless a further new one is introduced). Focus spaces are incremental in nature, with the information in spaces 'lower' in the stack accessible to those at higher levels, but not vice versa. The idea is that, at the end of the discourse, the stack of focus spaces will be empty, because all the topics in the discourse will have been exhausted.

The attentional state is controlled by the DSPs; it is the relations between the DSPs that determine the pushing and popping of focus spaces. For example, the cue phrase 'but anyway' can be used to pop the old focus space off the stack, indicate the satisfactory conclusion of the current DSP, and resume a prior DSP.

The main point of depicting attentional state in terms of focus spaces is to constrain the range of information under consideration at any one time when processing an utterance in a discourse. Lochbaum et al. (1994) and Lochbaum et al. (2000) draw a parallel between the structure of a computer program and the structure of discourse. In a computer program, lines of code are grouped together into procedures and functions that have a single purpose within the program

structure. Every line of code in a function contributes to the overall purpose of the function; this is analogous to a discourse segment in which each utterance contributes to the DSP. The function itself contributes to the overall purpose of the program, just as discourse segments contribute to the DP. So the purpose of the functions and their relationships with each other, and to the program as a whole, are directly comparable to the intentional structure of a discourse.

Similarly, the variables and constants used in a program are governed by scoping rules and are only successfully used within the functions in which they have been defined. The entities within such chunks of code are not 'visible' to, or referable by, code in the rest of the program. If a variable has the same name in different sections of code, under different scopes (or 'focuses'), then they are in fact referring to different entities. An example of this in discourse is the use of indexical referring words, such as 'him', which might refer at different stages of the discourse to Tom or Dick, depending on the context. So the scoping rules that determine the values of entities within a program's structure are directly comparable to the attentional structure of a discourse.

Grosz and Sidner provide thorough and convincing arguments for their theory of discourse structure. Although it is developed with task-oriented discourse in mind, they claim that this framework encompasses a more general approach to discourse than previous attempts. They claim that the three elements of discourse structure are necessary in order to simplify and extend the explanation of occurrences of linguistic phenomena such as the resolution of referring expressions, interruptions and cue words and phrases.

Although the model Grosz and Sidner present has a number of appealing characteristics and goes some way towards explaining the functional structure of language, there are also some gaps that are not accounted for. The theory they develop explains discourse coherence from the point of view of topic and focus, which is also linked to related models of intention and linguistic structure. However, the focus space is determined by syntactic factors, mapped to by the use of conversational cues; there is no inclusion of the semantic factors of the content of utterances in the process they portray, nor any explanation of how the meaning relates to the intentional structure or the attentional state. There is no account of how the underlying purpose can be recognised merely from the structure, nor how, once recognised, it influences the attentional state of the discourse.

The work of Grosz and Sidner deals with some fundamental issues concerning discourse that I hope to be able to tie into the presentation of my own model (see Chapters 9 and 10 for details). For example, they make a distinction between the initiating conversational participant (ICP) and the other conversational participant (OCP), which are usually termed 'speaker' and 'hearer' respectively in traditional approaches. Although I am not convinced that the fact of initiation is

in itself important, the fact that a distinction is made at all is a concession to the different roles of speaker and hearer that is unusual in the discourse literature. Grosz and Sidner go on to develop the notion of *Shared Plans* between participants, and it is at this point that I part company with them theoretically. As I have argued already, it is difficult to uphold the psychological validity of the claim that we have mutually identical beliefs and shared plans. The structural approach they present however, when coupled with some account of utterance level intention, may provide the beginnings of a powerful model for tracking intention in conversation. Although having no apparently direct bearing on speech act theory, Grosz and Sidner's work has been influential in the development of other structure-based speech act systems, which I shall be describing in the rest of this section.

There are two approaches to dialogue interpretation that are related to discourse structure theory and are generally viewed as being in contrast to the intention-based paradigm reviewed in Section 6.1. One, conversational game theory, is related the linguistic structural theories of conversation (as discussed in Chapter 5); the other, context change theory, views utterances (and their associated speech act interpretation(s)) as functions which update the context within which the dialogue occurs. It is these two theories that, in my opinion, hold out the best hope for finding a generic model for conversational structures, and therefore also for effective computational implementation. I have used an adaptation of some aspects of both theories in the development of my own model (which I shall present in Chapters 8 and 9).

6.2.2 Conversational Game Theory

Thinking about conversation as a kind of 'game' is a useful metaphor for the kinds of interaction patterns that occur within discourse. These conversational games correspond very closely to the nested levels proposed by Sinclair and Coulthard (1975) in their analysis of classroom interactions, and show similar structure to Grosz and Sidner's (1986) discourse segments.

At the highest level, dialogues are made up of **transactions**. Each transaction achieves some important sub-goal that contributes to the overall task. The size and composition of a transaction is dependent upon the type of task being performed. (These are analogous to the DSPs of Grosz and Sidner.)

Transactions are made up of **conversational games**. These are sometimes also called 'dialogue games' (Power 1979, Carlson 1983 and Mann 1988), or 'interactions' (Houghton 1986), or 'exchanges' (Sinclair and Coulthard 1975). Games have different purposes depending upon whether they are for example information-seeking or information-providing, initiating or responding. The expectations on the participants are different depending upon their role. Games are themselves composed of **conversational moves**, which are either initiation or

response moves (much like adjacency pairs, as described in Section 5.1); it is the moves that correspond to dialogue acts. A game consists of the initiative move and all the utterances that follow it. Games can nest within each other, for example when a participant needs extra information before being able to answer a question:

A: Do you want to go to the cinema on Saturday?
B: What time?
A: One o'clock in the afternoon.
B: I can't, I'm going out to lunch with Fred.

So the structure of a dialogue can be shown in much the same way as Sinclair and Coulthard's structure was represented in Figure 5.5. The main difference is that at the move level, one can substitute a (related) game instead of a response. I have tried to represent this diagrammatically in Figure 6.4.

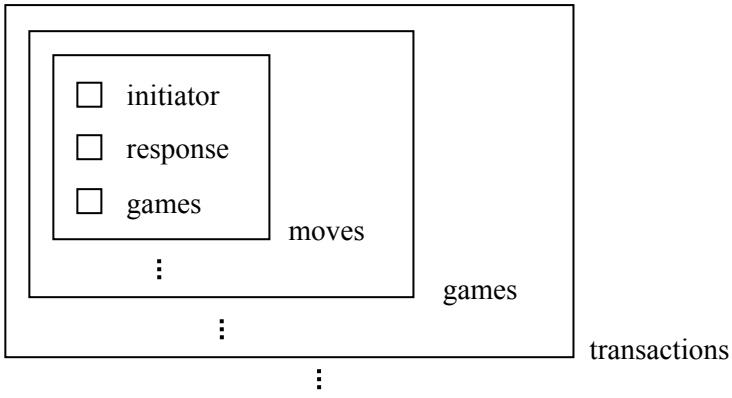


Figure 6.4 Conversational game structure.

Mann (1988) lists the following information that typically defines a game:

(i) There are different roles for participants (initiator and responder) depending upon the state of the game.

(ii) The game has an illocutionary point.

(iii) The responder has goals (or commitments) upon accepting the game.

(iv) There are a number of constraints that must be met for the correct performance of a game[7]:

  (a) The initiator and responder must pursue their respective goals.

  (b) Goals must be believed to be feasible.

  (c) The illocutionary point must not already be achieved.

---

[7] These are very similar to Searle's conditions for the performance of speech acts (see Section 3.2).

(d) The initiator must have the right to initiate the game.

(e) The responder must be willing to pursue the game's proposed goals upon accepting the game.

Game	Illocutionary point	Goals of *R*	Conventional Conditions
information-seeking	*I* knows *Q*	*I* knows *Q*	*R* knows *Q*
information-offering	*R* knows *P*	*R* knows *P*	*I* knows *P*; *R*'s information and *P* are consistent
information-probing	*I* knows whether *R* knows *Q*	*R* informs *I* of *R*'s knowledge about *Q*	*I* knows *Q*
helping	*I* is able to perform *A*	*I* is able to perform *A*	*R* is able to cause *I* to be able to perform *A*; *I* has the right to perform *A*
dispute	*R* believes *P*	*R* justifies that *I* might not believe *P*	*I* believes *P*; *R* does not believe *P*
permission-seeking	*I* knows that *R* grants the right that I perform *A*	*R* grants the right that *I* performs *A* or not, and *I* knows this	*I* wants to perform *A*; *I* does not have the right to perform *A*; *R* can grant the right to perform *A*
action-seeking	*R* causes *A* to be performed	*R* causes *A* to be performed	*R* would not cause *A* to be performed in the normal course of events

Table 6.1 Examples of Mann's dialogue games.

In Table 6.1, some of the game types suggested by Mann (1988: 515) are reproduced. This list is by no means complete. To interpret Table 6.1, one must know the following: *I* is the initiator, *R* is the responder, *Q* is the content of a question, *P* is the propositional content of the utterance, and *A* is the action proposed by the initiator. The illocutionary point is the goal that the speaker is trying to achieve by the move he has just made (e.g. the point of an information-seeking game is that *I* should get to know the answer to *Q*). The goals of *R* are those to which he must adopt if he is to co-operate with *I* (e.g. the goals that *R* must adopt in an information-seeking game is that *I* should get to know the answer to *Q*). The conventional conditions are those that must obtain in order for the illocutionary point of a move to succeed (e.g. the conventional conditions of an information-seeking game is that *R* knows the answer to *Q*).

Conversational game theory formalises the link between an utterance type and a response type by suggesting that a fundamental unit of discourse encompasses more than one utterance or move. This allows for the important role that both initiator (addressor) and responder (addressee) have in a dialogue. For example an information-seeking game not only consists of a yes-no question, but also the yes-no reply.

The conversational games developed by Kowtko et al. (1992) and Carletta et al. (1997) were based on their studies of the MAPTASK corpus. The MAPTASK dialogues were collected at the Human Communication Research Centre (HCRC) in Edinburgh. Dialogues were centred on a route-finding task. Two participants would each be given a map with similar, though not

identical, features marked on them. One of the subjects has a route drawn on their map, which the other subject must reproduce on their map without seeing the original. Knowledge of the route is transferred orally between participants, who have therefore fixed roles in the dialogue – one of instruction-giver and one of instruction-follower.

The games that are postulated from the analysis and mark-up of the MAPTASK corpus are: INSTRUCT, EXPLAIN, CHECK, ALIGN, QUERY-YN, and QUERY-W. As I have mentioned above, there are two categories of move: *initiating* and *response*. The games are identified according to their type of initiating move:

Initiating Moves (which set up an expectation of response):

INSTRUCT: Commands (or instructs) the dialogue partner to perform some action.

EXPLAIN: States some information that is not elicited by the partner.

CHECK: Requests confirmation about information that the initiator has reason to believe is true, but is unsure about.

ALIGN: Checks the attention or agreement of the responder with regard to the current state of the game, or checks whether he is ready for the next move.

QUERY-YN: Asks a yes-no question (i.e. one that expects either the answer 'Yes' or 'No') for information that is unknown (because otherwise it would be a CHECK or an ALIGN move).

QUERY-W: Asks a question that expects a more complete answer than QUERY-YN. A question that expects the responder to choose from a set of answers (included in the question content) is also subsumed by this move.

Response Moves (which complete the current game):

ACKNOWLEDGEMENT: Shows (in a minimal way) that the responder has heard and understood the move to which it responds.

REPLY-Y: Responds affirmatively (usually in response to a QUERY-YN, ALIGN or CHECK move).

REPLY-N: Responds negatively (usually in response to a QUERY-YN, ALIGN or CHECK move).

REPLY-W: Responds with requested, elicited information.

CLARIFY: Clarifies or rephrases given information.

There is one more type of move that is made in preparation for a new game: the READY move. This move does not fall under either of the categories given above, but corresponds roughly with Schegloff's (1988) pre-sequences. READY moves are characteristically short phrases (such as 'Okay', 'Right' and 'Now') to warn the responder that a new game is beginning. The decision tree used for the identification of different moves is given in Figure 6.5 (taken from Carletta et al. 1997).

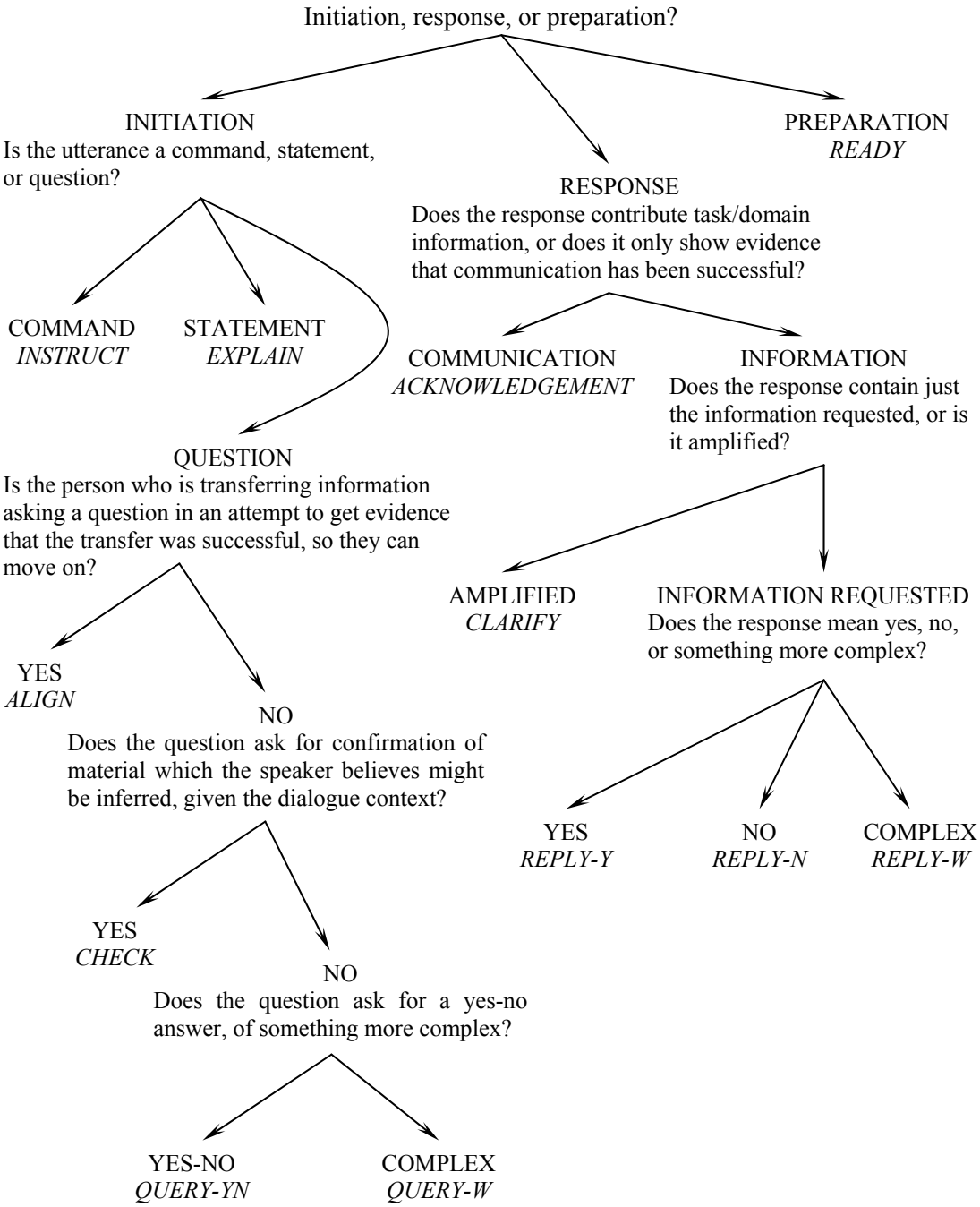


Figure 6.5 Conversational move categories for MAPTASK.

Conversational game theory provides a useful framework for analysing task-oriented dialogues. However, as a theoretical explanation of dialogue, it remains quite weak. Firstly, it is difficult

to see how the ascription of mental states from an utterance differs in many respects from intention-based frameworks, except that some of the patterns of inference are hard-wired rather than derived from first principles. If one compares the kinds of dialogue acts covered in the MAPTASK corpus by Carletta et al. (1997), with those in the TRAINS dialogue system by Allen et al. (1994), there is a considerable overlap.

Secondly, conversational games suffer from a lack of constraint similar to that of Grosz and Sidner's theory. For example, it is unclear that one can specify exactly how many different types of game there are. The game types defined within MAPTASK are arguably quite coarse-grained in nature; however other projects have developed much finer-grained 'games'. The VERBMOBIL system (which I shall be covering in Section 6.2.5 and also again briefly in 7.3.2) uses specific 'games' such as INTRODUCE_NAME and REQUEST_SUGGEST_DATE. Although these are developed within a certain domain from abstract classes of games, one can see that defining games at this level would mean an endless proliferation of types.

Thirdly, what exactly distinguishes a conversational *game* from a *move*? One cannot identify a type of game from a standardised sequence of moves (as has been suggested by Houghton 1986). One would like to be able to say that for example a QUERY-W move/game is completed by a REPLY-W move. However, in some cases, the QUERY-W seems to span a larger section of dialogue than simply the REPLY-W, as one can see from the example given below (adapted from Pulman 1996):

		MOVE	***GAME***
(1)	*I*:	Where would you like to go?	QUERY-W
(2)	*R*:	Edwinstowe.	REPLY-W
(3)	*I*:	Edwinstowe?	CHECK
(4)	*R*:	Yes.	CLARIFY
(5)	*I*:	Please wait.	ALIGN/ACKNOWLEDGE
(6)	*I*:	Is that Edwinstowe in Nottingham?	QUERY-YN/CHECK
(7)	*R*:	Yes.	REPLY-Y/CLARIFY

Lewin et al. (1993) used the MAPTASK move scheme in a computational model to help predict what move the user is trying to make in a route enquiry system. They suggest that one could define a game as the sequence of dialogue acts (moves) that actively change the status of the propositions that are the focus of the current talk exchange from 'proposals' to 'agreed commitment'. This does go some way towards explaining dialogues like the one given above; although the context is changed after every move, and three separate conversational games are played out, there is only one significant change to the agreed commitments of the participants. Moves (3) – (7) are checks to make sure that the information in move (2) is grounded correctly and that both players are talking about the same place. Lewin et al. also point out that the

second and third games in this dialogue are possibly not planned at all, but are generated dynamically in the current conversational context.

There are still further problems that are highlighted by the above dialogue. It is often very difficult to assign a move type consistently to an utterance. In moves (6) and (7) we could choose from two pairs of moves. This suggests that there is a problem with the level at which moves are ascribed to utterances, or that some utterances are used for multiple purposes. Conversational game theory has nothing to say about this phenomenon. We can also notice from move (5) that some games are only one move long, or are in actual fact completed by a non-linguistic action – in move (5) it is the act of waiting. Again, this is unaccounted for.

The conversational game theory framework covers those dialogue acts that do not fit very well into traditional speech act theory, such as elements of dialogue control and communicating the state of the game (such as acts of alignment, which check that both participants are at the same point in the dialogue). Arguably this conflates two different phenomena into one theoretical account.

Bunt (1994) refines these dialogue control acts in his dynamic interpretation theory (implemented within the theory of context change). He distinguishes between different types of context, and defines different dialogue acts for each type. It is within the theory of context change that I have developed my own model of speech acts – I shall describe both the theory and the model in greater detail in Chapters 8 and 9. As far as I am aware, dynamic interpretation theory is the only other attempt to formalise and implement computationally a speech act recognition system based on context change.

6.2.3 Dynamic Interpretation Theory

Bunt (1989, 1994, 1995, 2000 and 2001) specifies a range of dialogue acts as part of his theory based on context change. Utterances are used as a means of updating the context, via the recognition of the dialogue act being performed. Dialogue acts are the units of function that the speaker employs in order to change the context, and do not necessarily have a one-to-one correspondence with an utterance. An utterance can perform one or more dialogue acts at a time. A dialogue act is composed of the semantic content of an utterance, as well as a communicative function that tells us what is the speaker's intended use of the semantic content within the context (much like Searle's representation of speech act theory).

Bunt's dialogue act classification is developed from task-oriented discourse. He develops his dialogue act taxonomy by the analysis of airport reservation dialogues. The taxonomy of dialogue acts covers the types of phenomena found in such task-based dialogues and is not claimed to be general or complete for all different types of spoken interaction.

Dynamic interpretation theory (DIT) defines an utterance's meanings in terms of how it affects the context of the current dialogue. The word 'context' has been used in the literature to stand for many different types of thing: the physical environment, preceding discourse, or the domain of the discourse (or topic). What these all have in common is that they refer to factors that have a bearing upon the interpretation of human communicative behaviour. Bunt distinguishes five different types of context in communication: *linguistic*, *semantic*, *cognitive*, *physical* (and *perceptual*), and *social*. Furthermore, for each of these types, he distinguishes between *global* aspects (that are established at the beginning of the dialogue and are constant throughout), and *local* aspects (that have values that change and develop as the interaction proceeds). The local aspects are in essence the focus and topic shift aspects of DIT and are significant for the ongoing continuation of a dialogue. A more detailed list of the different types of context is given below:

(1) **Linguistic context**: The surrounding linguistic material, 'raw' material, as well as relevant properties, determined by analysis (the dialogue history).

(2) **Semantic context**: The current state of the underlying tasks: properties and facts of the task domain.

(3) **Cognitive context**: The participants' states of processing and models of each other's states.

(4) **Physical and perceptual context**: The availability of communicative and perceptual channels; partner's presence and attention.

(5) **Social context**: The communicative rights, obligations and constraints of each participant. (Adapted from Bunt 2000: 100)

Unlike conversational game theory, dynamic interpretation theory distinguishes between task-related acts and dialogue control acts. The former are those acts that are more traditionally associated with illocutionary acts, while the latter are concerned with controlling the flow of the interaction, e.g. elaborations, paraphrasing, repetitions, acknowledgements. All dialogue acts will have an effect upon the cognitive context when interpreted by the hearer. The difference between task-oriented acts and dialogue control acts is that the former will change the *semantic context*, while the latter will not (although it might change the *physical* or *social contexts*).

Task-Oriented Acts: Bunt identifies acts that are used by the speaker to achieve domain-related goals, as well as purely communicative ones. These task-oriented acts are further sub-classified as either information-seeking or information-providing. The full set of task-oriented acts can be seen in Figure 6.6. Labels in lowercase show the classes of communicative function, while the uppercase ones show the communicative functions themselves.

Task-oriented acts are defined by the mental attitudes (of the speaker and hearer) that must hold true for their successful performance in the current context. This is very similar to other approaches (intention-based for example), as we can see from the preconditions defined for an INFORM act:

BELIEVES (s, prop),
WANTS (s, BELIEVES (h, prop))

Acts are specified in the tree structure as shown in Figure 6.6; the nodes at the end of the tree inherit, or strengthen, the preconditions of the parent nodes. So for example, weak-INFORM is the more general dialogue act type than INFORM proper, and CORRECTION.

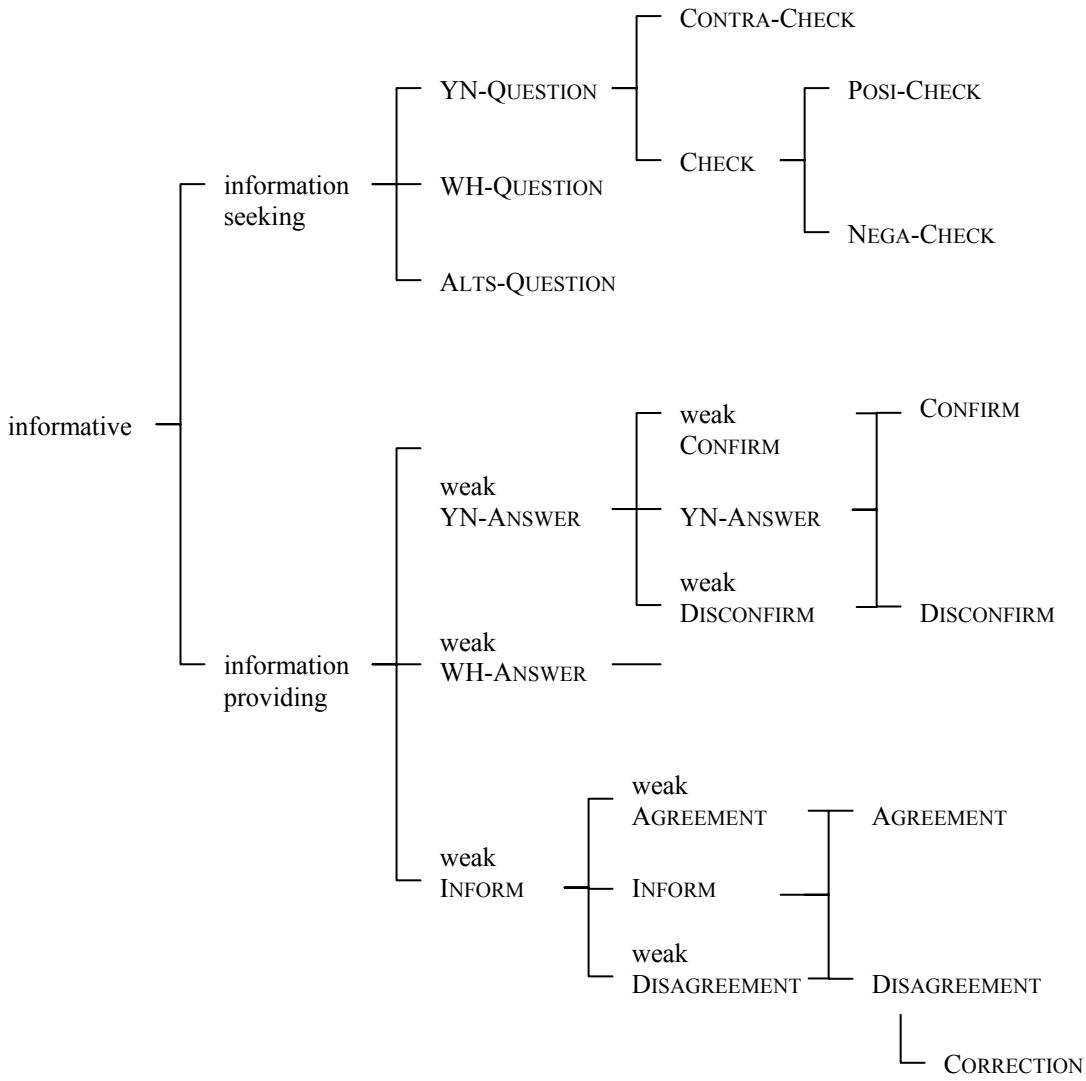


Figure 6.6 Task-oriented communicative functions.

Figure 6.7 shows the gradual incrementation of preconditions between weak-INFORM and CORRECTION. Notice that in addition to BELIEVES and WANTS, Bunt uses a SUSPECTS operator

to indicate weak belief. So, DISAGREEMENT inherits the conditions from weak-DISAGREEMENT, but converts the suspicion into belief.

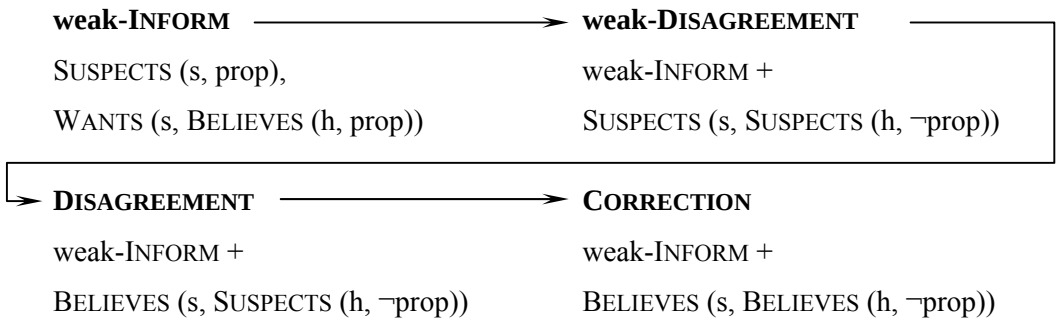


Figure 6.7 Precondition inheritance from weak- INFORM.

While task-oriented acts aim to contribute to the achievement of a communicative goal motivated by the task in question, dialogue control acts aim to satisfy a communicative goal motivated by the state of the interaction and reflect a form of social behaviour.

Dialogue Control Acts: Dialogue control acts are further subdivided by Bunt into three categories: feedback acts, interaction management acts and social obligation management acts. A breakdown of these acts is shown in Figure 6.8 (from Bunt 1995).

Feedback acts inform the hearer of the speaker's understanding of the hearer's last dialogue act. Feedback is positive in the case when the speaker believes he has interpreted the hearer's utterance correctly, and negative when he has encountered some problem. There are also two kinds of feedback: auto- and allo-feedback. Auto-feedback refers to the speaker's self-analysis of his understanding of what the hearer has just said (2a). Allo-feedback is the speaker's analysis of the hearer's displayed understanding (or misunderstanding) of a speaker's earlier utterance (3b). E.g.:

(1) A: Could you give me the departure times for Munich next Tuesday morning please?

(2a) B: Munich, Tuesday morning

(3a) A: Correct

(2b) B: Munich, Tuesday evening

(3b) A: No, Tuesday *morning*

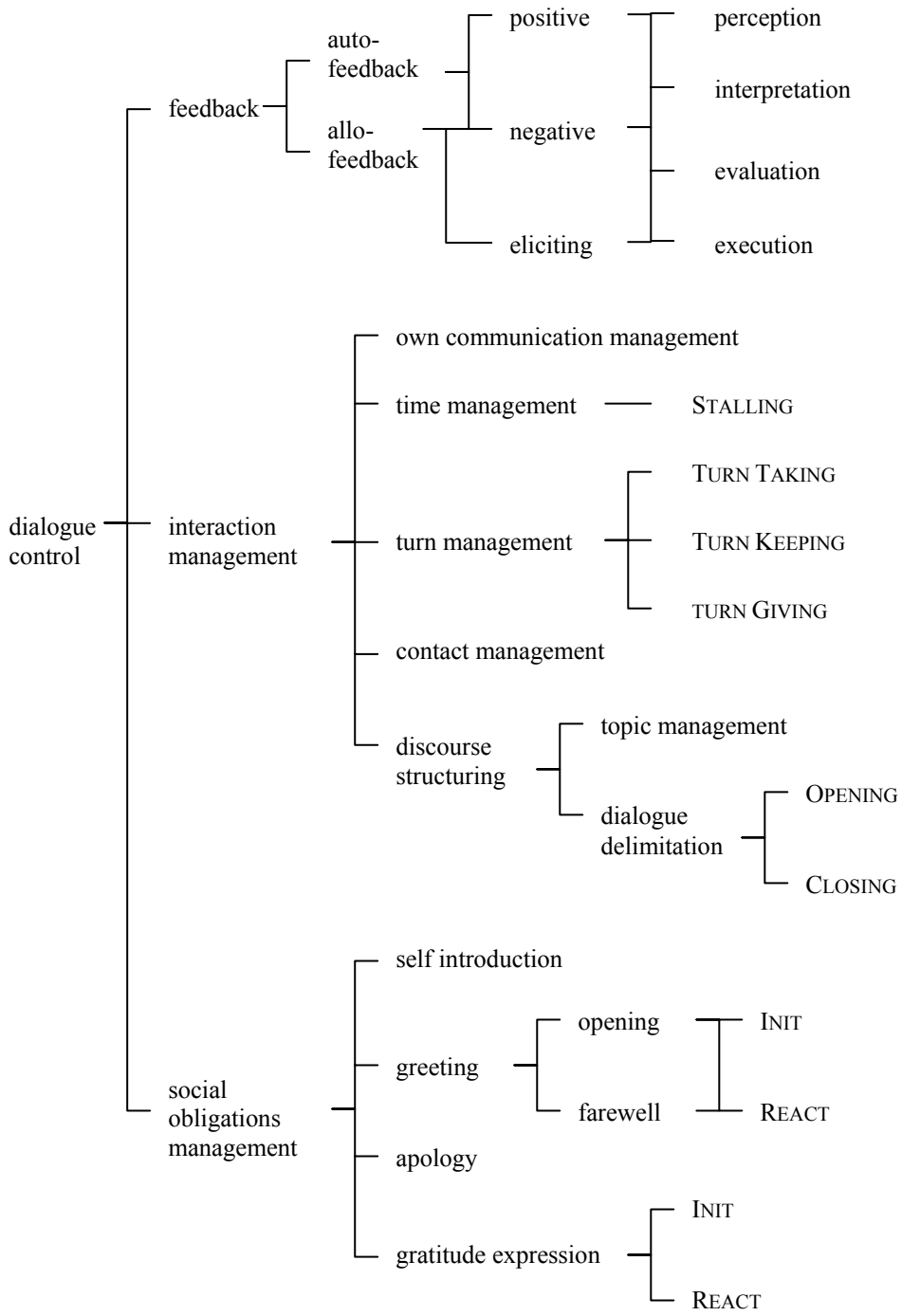


Figure 6.8 Dialogue control functions.

Interaction management acts are acts that deal with and control the flow and structure of the dialogue. The important factors of interaction management are as follows:

Own communication management (self-correction, 'Er… I mean')
Time management (stalling, 'Hold on a minute')
Turn management (turn-taking, -keeping and -giving)
Contact management (attention grabbing utterances, eye contact, etc.)
Discourse structuring (topic management, dialogue delimitation – opening and closing)

Social obligation management acts are concerned with the control of entering, maintaining and leaving a social interaction with a dialogue participant. Such social aspects of communication are present even in task-oriented dialogues and so must be included for a dialogue system to be robust enough to deal with real interactions. Social obligation acts are performed by a closed class of utterance types that are generally conventional in nature. These types of utterances have the property of placing what Bunt terms a *reactive pressure*[8] on the co-participant to respond using a corresponding related prescribed utterance type. For example, if one participant says 'Thank you', the addressee is placed under some social obligation to acknowledge the gratitude expressed by the speaker in some conventional way, perhaps 'You're welcome', or 'No problem'. Bunt identifies five frequent situations in which social obligations use this closed-class set of formulations:

1. Self-introduction.
2. Hello greeting at the beginning of a dialogue.
3. Farewell greeting at the end of a dialogue.
4. Apology.
5. Expression of gratitude.

These all come in initiative and reactive pairs, so that a self-introduction and both types of greeting will be reciprocated, an apology will (usually) be accepted, and an expression of gratitude will be followed by an expression of deprecation.

Both types of dialogue act, task-oriented and dialogue control, are recognised by the definition of their preconditions. Instead of attaching effects to the dialogue act, Bunt has specified separate rules for updating the cognitive contexts of the speaker and the hearer of a certain dialogue act. The effect for a speaker is:

> For all preconditions *c* in the dialogue act, add:
> MB (s, h, SUSPECTS (s, BELIEVES (h, *c*)))

This means that after the speaker has performed the dialogue act, a mutual belief (MB) is established that the speaker SUSPECTS that the hearer BELIEVES every precondition for the satisfaction of the dialogue act. The effect for a hearer however is:

---

[8] Bunt (1995) points out that this is a similar idea to others found in the literature: Allwood's (1994) *communicative obligations*, Schegloff and Sacks's (1973) *adjacency pairs*, and Levinson's (1983) *preference organisation*.

For all preconditions *c* in the dialogue act, add:

BELIEVES (h, *c*),

MB (h, s, SUSPECTS (s, BELIEVES (h, *c*)))

After every dialogue act performed by the speaker, the hearer BELIEVES all the preconditions of that act are true, as well as MB that the speaker SUSPECTS that the hearer BELIEVES all the conditions.

These rewrite rules are slightly odd for at least a couple of reasons. If a hearer automatically assumes that all the preconditions of a dialogue act are true, why is it that the speaker only SUSPECTS as much? Also, because the hearer's effects cause him naïvely to believe the preconditions of every dialogue act, this leaves no room for the possibility of the speaker either being mistaken, or intentionally attempting to deceive the hearer in some way. Clearly it is psychologically unsound to make such an assumption; a hearer does not automatically believe all he is told. Bunt does not describe how his system would recover from the hearer's misinterpretation of a dialogue act, or from his assumption of incorrect beliefs.

Aside from these problems with the way the rules for the effects of a dialogue act have been handled and formulated, DIT presents a very appealing approach for the derivation of dialogue act effects on the context. Acts are defined and identified by their preconditions rather than their effects, which is intuitively an attractive idea.

Also appealing is the way that utterances are treated as multi-functional entities. There is no one-to-one correspondence between the utterance and the dialogue act, or acts, that it can perform. An utterance may have more than one meaning, for any one of the following reasons:

(1) *Indirectness*: As I have already discussed in Chapter 4.

(2) *Functional subsumption*: Some dialogue acts subsume others, so for example, the utterance of 'I will come over tonight' is not just a promise, but also an informative statement.

(3) *Functional multi-dimensionality*: Elements of task-oriented functions are commonly combined with elements of dialogue control in one utterance. Similarly, different elements of dialogue control can combine together in one utterance. So for example, 'Thank you' may not only be an expression of gratitude, but also may provide feedback for a previous utterance by showing that the hearer's utterance was heard and understood. Depending upon the kind of intonation, this will also indicate something about who is to take the next turn.

As far as I know, DIT represents the first system to derive the effects of dialogue acts on the context, for the purpose of task-oriented interaction, from a principled basis. I shall be returning to the theory of context change in Chapters 8 and 9, when I shall attempt to generalise and formalise the theory for casual conversation.

Most of the approaches to speech act recognition covered so far have been dependent on the application of logical inferences to work out what action a person intends to perform in the production of an utterance and how this affects the state of their beliefs. In essence, these are theories of mental models of belief, goals and intention, based on Gricean ideas of meaning interpretation.

6.2.4 TRINDI

Before looking at how statistical methods have been used to extract dialogue patterns in task-oriented and general conversation in Section 6.2.5, I will briefly cover the TRINDI (2001) project, which was based on the work carried out in the various research projects covered so far in this chapter.

TRINDI (which stands for **T**ask-o**R**iented **IN**structional **DI**alogue) was one attempt to unify plan- and structure-based theories under one over-arching structure. The TRINDI project aimed to provide a tool for the development of moderately complex systems with the ability to reason, which would also be easily 'portable' from one domain to another, and from one language to another. It purported to provide a generic framework based on informational states and update rules, which would be flexible enough to model the varying different dialogue management systems that are already available, while increasing the potential for a more functional system in the future. The resulting toolkit that was developed was applied to:

- The Gothenburg Dialogue System, building on the work of Ginzburg (1996) and Cooper and Larsson (1999) on Questions Under Discussion.
- The dialogue model of Poesio and Traum (1997, 1998).
- A prototype system based on DRT (Kamp and Reyle 1993) called MIDAS (Multiple Inference-based Dialogue Analysis System) by Bos and Gabsdil (2000).
- The SRI Autoroute Demonstrator, based on conversational game theory (Power 1979, Houghton 1986, Carletta 1997 et al.) and recast using the TRINDI architecture.

All of these toy systems are built to show the applicability of the TRINDI architecture for systems coming from different research traditions. One small criticism is that no evaluation is given to show whether the performance of the original systems and that of the ones developed using the TRINDI toolkit are comparable. I believe that this was because the aim of the application of the TRINDIKIT was to test the finite state techniques it employed and explore its weaknesses rather than to present a finished product that could instantly replace other systems and methods of dialogue system development.

The project also devised a 'best practice' tick-list against which it evaluated some of the leading dialogue management systems software available: (1) software for building spoken dialogue systems: Nuance Communications Dialog Builder (used commercially for building automated enquiry systems) and the CSLU Speech Toolkit (for rapid development of spoken language systems, primarily for teaching purposes); and (2) actual dialogue systems: the Philips Train Timetable System, the SRI Autoroute System, TRAINS and VERBMOBIL.

TRINDI was an ambitious project that held out a promising means of combining different approaches for the development of task-oriented dialogue systems. It is a shame therefore that there appears to have been little uptake of the work and research in the years that followed its release.

6.2.5 Statistical Approaches

In the last few decades there has been a trend towards using stochastic methods in linguistics for processing large bodies of text called *corpora* in order to extract features of language automatically. These have been surprisingly successful for predicting and assigning grammatical (e.g. parts of speech) labels, with high accuracy rates.

These statistical approaches have also been applied to the recognition of speech acts with varying degrees of success. Work has concentrated on the derivation of models of recognition by empirical methods: either by trying to predict the speech act from the constituent parts of an utterance, or from the pattern of recognised speech acts from previous utterances in the dialogue, or from a mixture of both processes. The major research in these statistically-based models has been carried out in Germany and Japan, for applications in machine translation, and in North America with the Discourse Research Initiative (DRI) where computational speech recognition has been the main aim (for the automation of call response for information phone lines for instance).

I wish to review two of the most influential projects in particular: the VERBMOBIL system and work on the SWITCHBOARD[9] corpus with DAMSL[10]-based, 'flattened' dialogue act[11] annotation

---

9 I shall describe the composition and collection of this corpus in Section 7.2.1.

10 Which stands for **D**ialogue **A**ct **M**arkup in **S**everal **L**ayers.

11 In the discussion that follows I shall refer to *dialogue* acts when talking about other annotation schemes in order to distinguish these from what I mean by the term *speech* act. I wish to draw a theoretical line between my definition, and the large number of different types and levels of language functions that are generally accepted under the heading of dialogue act. I have retained the original terminology of *speech* act when referring to illocutionary acts to emphasise that it is only the act as it occurs naturally in conversation that is primarily of interest to my research.

scheme. I shall be mentioning other work also, but am interested in these two especially because they are representative of the rest, and also stand for two different methodologies in dialogue act recognition. I shall be reviewing the dialogue act schemes themselves critically in Chapter 7. Here I am primarily concerned with describing the successes and failures of the application of the schemes. In a sense perhaps it could be argued that an appraisal of the schemes should come before that of the application, as the results are in many ways prejudiced by problems with the schemes. Still, it is important to set the scene before showing why current approaches are insufficient for the recognition of speech acts in conversation.

6.2.5.1 VERBMOBIL

The European based project VERBMOBIL (Alexandersson et al. 1998) was set up primarily with a view to advancing machine translation. The use of dialogue act recognition here is mainly to provide a framework for translation invariance. The idea being that if you can recognise the act performed in one language, along with the various parameters associated with its instantiation in an utterance, and you also know how to construct the same dialogue act in the target language, with the relevant information plugged in appropriately, then this might provide an aid to quick and efficient translation.

In the first phase of VERBMOBIL, the scenario was restricted to the scheduling and negotiation of appointments. The second phase was extended to deal with travel planning, a significantly more complicated domain. Although both phases cover information-seeking and information-giving dialogues dealing with mixed initiative interactions, the second edition of VERBMOBIL is not just concerned with negotiating a single task, but must deal with various kinds of user input. In the first phase of the project, input utterances were gathered from two participants who would say something, then press a button to indicate that the process of translation should go ahead and so on. In the latest version, the system automatically monitors input and translates without the need to press buttons. It is interesting to note that the dialogue acts have been updated too in VERBMOBIL-2: interesting because to a certain extent it demonstrates an approach to dialogue act definition that lacks rigour and is not intended to be applicable generically, but to be redesigned or revised when used in different scenarios. The model has specifically been structured to work well in one area, one task.

In VERBMOBIL, dialogue acts express the speaker's *primary communicative intention* that is intended to be conveyed by the utterance or discourse segment. The recognition of the dialogue act being performed is important for (at least) five reasons:

- **transfer**: The primary aim of dialogue act recognition in VERBMOBIL (as mentioned above) is to produce an appropriate translation of an utterance in one language into an utterance in another language. The identification of the dialogue act could help to disambiguate in the

case when an utterance might have more than one translation, so that the best option is chosen.

- **shallow processing**: In VERBMOBIL two varieties of processing are exploited in order to extract and assign meaning. Deep processing uses knowledge-based techniques, whereas shallow processing uses surface grammatical and statistical algorithms. At the shallow processing stage, dialogue acts are used to determine the selection of 'template' that will generate the equivalent utterance in the target language.

- **dialogue memory**: Dialogue acts provide the means of keeping track of where participants are in a dialogue (e.g. whether a date for an appointment has been accepted or rejected yet), and for updating the focus of the dialogue.

- **prosody**: In spontaneous speech, even in a constrained dialogue scenario, the flow of output is often fragmented or faulty due to hearers rephrasing their utterances or self-correcting mid utterance. It is unrealistic therefore for complete sentences to be used as the most fundamental unit of processing. The basic item that is processed is the smallest segment that can be tagged with a dialogue act. The system uses a statistical model of prosodic contours as matched with typical associated dialogue acts, which is trained on transcribed dialogue annotated with both prosodic and dialogue act information.

- **summary generation**: It is the recognition of the underlying dialogue acts that allows the automatic summarisation of the dialogue between participants.

The VERBMOBIL dialogue acts have been tested and refined against the ability of human annotators to assign the same dialogue acts to the same utterances. The system recognises English and German dialogue acts with rates of accuracy between 65% and 75%; variations are dependent on the type of training and test data provided. The κ (kappa) value[12] of the difference between human and automated dialogue act recognition is between 0.61 and 0.65. It

---

[12] The kappa coefficient κ is used as a measure of inter-annotator agreement for a coding scheme (so that one can have some indication of the reliability of the coding scheme). The κ statistic is worked out by counting the number of times pairwise coders agree on the annotation of a feature with a category label (in this case, the labelling of an utterance with the same dialogue act). This then gives us an agreement percentage for the whole corpus. However, one must also take into account the agreement that might occur purely by chance, so the chance proportion of agreement is included in the equation to give κ:

$$\kappa = (P(A) - P(E)) / (1 - P(E))$$

where P(A) is the actual agreement, and P(E) expected agreement by chance. Coding schemes with overall reliabilities of $\kappa = 0.8$ or higher are good enough so that it is not necessary to try to improve on them; values between 0.67 and 0.8 allow tentative conclusions to be drawn but indicate that the scheme could be improved (Carletta 1996).

is the conventional and the 'regular' acts that are most readily identified correctly in dialogue, while the 'irregular' acts (such as DIGRESS) are the ones that most often cause the annotators difficulties, when the utterance fails to conform with the expectations set up by the rest of the dialogue. In other cases, the distinction between two different acts is blurred, and this also causes 'misclassification' (e.g. GIVE_REASON is often confused with SUGGEST or REJECT – this problem was solved in part by the introduction of an extra act, EXPLAINED_REJECT, which is the child of both GIVE_REASON and REJECT).

Although results are predictably good for conventional acts (of greeting, thanking, taking leave, feedback production, etc., which have fixed formulations in the lexicon and are therefore typically fairly easy to spot) and core acts (like SUGGEST, ACCEPT and REJECT) where the use of set words and phrases strongly influences the recognition of the dialogue act being performed, the statistical model does not work very well over all (see Reithinger and Klesen 1997).

So, recognition rates for the more obscure dialogue acts defined in VERBMOBIL are not very high, and this is despite the organisation of the choice of dialogue acts being constrained by a decision tree hierarchy. At each node of the tree the path one can take is decided by answering certain questions. The whole tree of dialogue acts is given in Figure 6.9 (from Alexandersson et al. 1998: 19).

An example of the kind of decision algorithm that is given with the tree is as follows:

> **Decision PROMOTE_TASK:**
>
> **if** the segment under consideration contains a request for information, where the content refers to an instance that has been already explicitly introduced in the previous discourse (i.e. which is not mentioned explicitly)
>
> **then** branch to REQUEST.
>
> **else if** the segment by the speaker to offer an action or concession on his/her part label with OFFER
>
> **else if** the speaker explicitly commits him-/herself to executing a specified action label with COMMIT
>
> **else if** the segment contains a suggestion which is given by means of an explicitly mentioned instance or an aspect of such an instance label with SUGGEST
>
> **else if** the segment contains a reaction to a previous part of the discourse **then** branch to FEEDBACK.
>
> **else** branch to INFORM. (Alexandersson et al. 1998: 23)

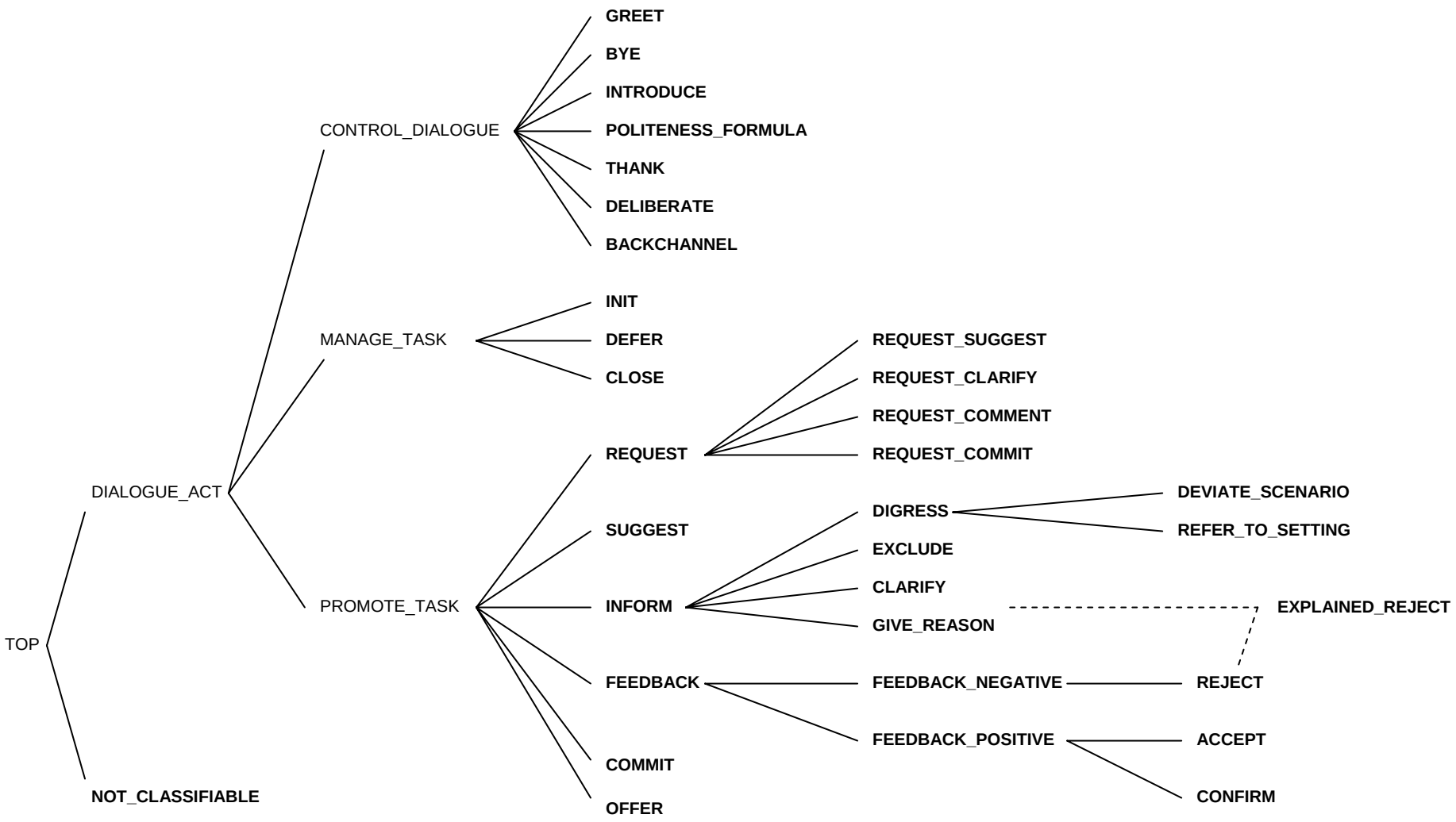


Figure 6.9 The VERBMOBIL dialogue act hierarchy.

An interesting mechanism that is implemented in VERBMOBIL is an extra layer of structure for dialogues. Dialogues are sub-divided into different phases. In negotiating dialogues, they distinguish between the following five phases (for a full description see Alexandersson el al. 1998: 10): **Hello**, **Opening**, **Negotiation**, **Closing** and **Good_Bye**. Alexandersson et al. (1998) claim that without the information about what stage the dialogue is in, it would not be possible to produce an accurate translation of some utterances. This claim appears to be backed up with evidence from the corpus collected. Some verb meanings are determined solely from their positioning within the structure of a conversation, i.e. whether we are in the opening, negotiating or closing stages will affect the interpretation of the verb. If this were the case, it would provide a very convincing argument for the identification of dialogue structure being a crucial component of an NLP system. It clearly shows that those analyses based on stochastic methods alone, are inadequate for the full understanding of language as used in conversation.

6.2.5.2 DAMSL and (SWITCHBOARD) SWBD-DAMSL

The DAMSL annotation scheme was influenced by the design of VERBMOBIL, and aimed at producing a generic, standard tag-set, from which specific dialogue act schemes could be developed for task-specific domains (Core 1998). This work is also related to work in plan recognition in that the development of the annotation schemes was strongly influenced by researchers in dialogue systems such as TRAINS (the simulated TRAINS dialogues were used to test the DAMSL scheme for instance). Allen, for example, was one of the two authors of the manual for annotation using the DAMSL scheme.

Although the DAMSL scheme has stemmed originally from an intention-based background, it also borrows heavily (and principally) from work in structure-based methods of analysis. These include: Schegloff et al.'s (1977) research on dialogue **repair**, Clark and Schaefer's (1989) work on **grounding**, as well as theories which categorise dialogue acts according to whether they are initiative, **forward-looking**, or responsive, **backward-looking** (Schegloff 1968, Schegloff and Sacks 1973, Schegloff 1988, Allwood et al. 1992 and Allwood 1995). These last features are also similar to the ideas proposed by researchers in conversational game theory.

So, DAMSL is in essence a general framework of task-related dialogue acts, some *forward-looking*, and some *backward-looking* (these are the acts that correspond most closely with traditional philosophical accounts of illocutionary acts by Austin and Searle). DAMSL also makes use of annotations that deal with the *communicative status* of an utterance (i.e. whether it is interpretable or not) and the *information-level* of an utterance (i.e. whether it is actually performing the task, or talking about the task, etc.).

An utterance has a forward-looking function if its production has an effect on the subsequent dialogue. The different types of tags for this category of dialogue act are as follows:

Forward-Looking-Function	
STATEMENT	a claim made by the speaker
ASSERT	a claim intended to be believed by hearer
REASSERT	a claim repeated by the speaker
OTHER	a comment made by the speaker
INFLUENCING-ADDRESSEE-FUTURE-ACTION	(equivalent to Searle's directives)
OPEN-OPTION	a weak suggestion or listing of options
ACTION-DIRECTIVE	an actual command
INFO-REQUEST	request for information
COMMITTING-SPEAKER-FUTURE-ACTION	(equivalent to Austin's commissives)
OFFER	speaker offers to do something
COMMIT	speaker commits to doing something
CONVENTIONAL	conventional formulations
OPENING	greetings
CLOSING	farewells
EXPLICIT-PERFORMATIVE	dialogue act named by the verb
EXCLAMATION	an exclamation (of surprise, etc.)
OTHER-FORWARD-FUNCTION	any other forward-looking function

Decision trees for some of the forward-looking dialogue acts that are most easy to confuse were given to the annotators to help code the different acts consistently. There are three – one each for STATEMENT, INFLUENCING-ADDRESSEE-FUTURE-ACTION and COMMITTING-SPEAKER-FUTURE-ACTION; these are shown in Figures 6.10 – 6.12.

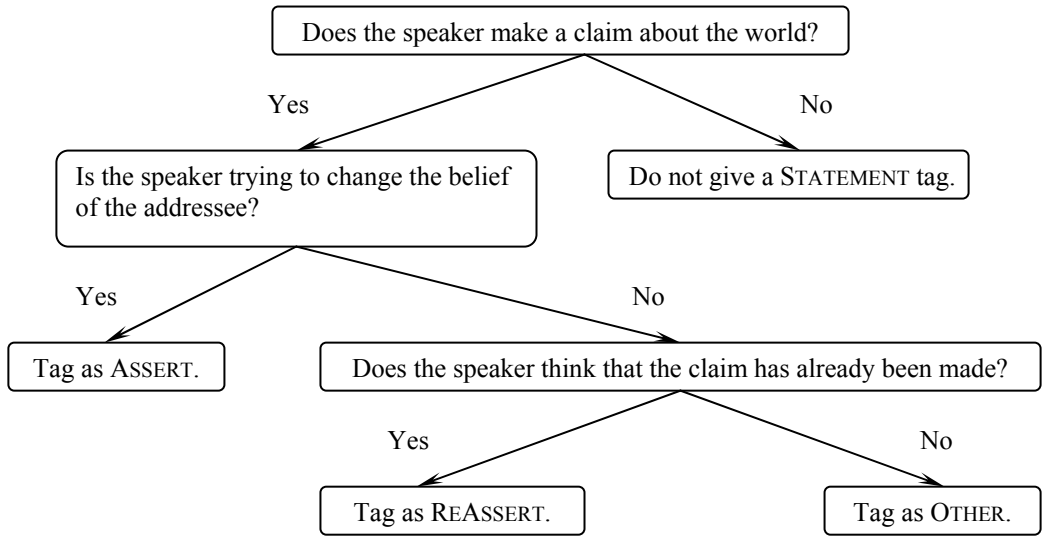


Figure 6.10 DAMSL decision tree for STATEMENT.

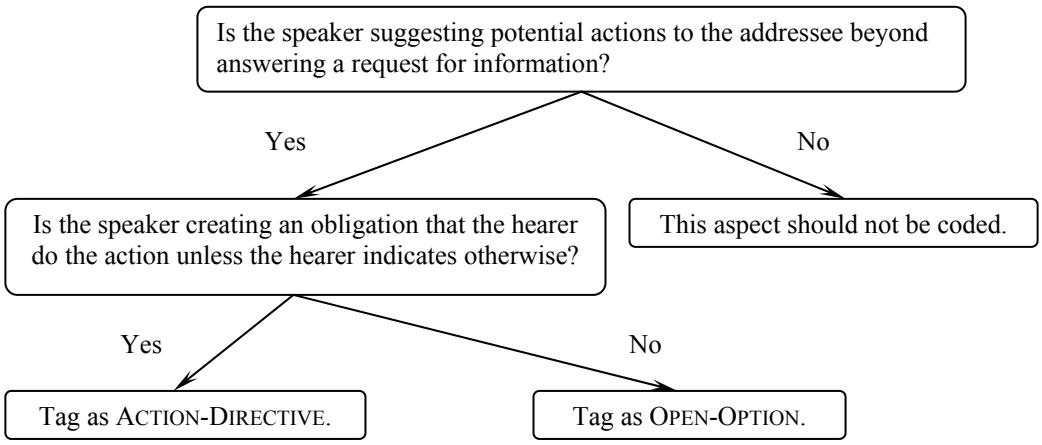


Figure 6.11 DAMSL decision tree for INFLUENCING-ADDRESSEE-FUTURE-ACTION.

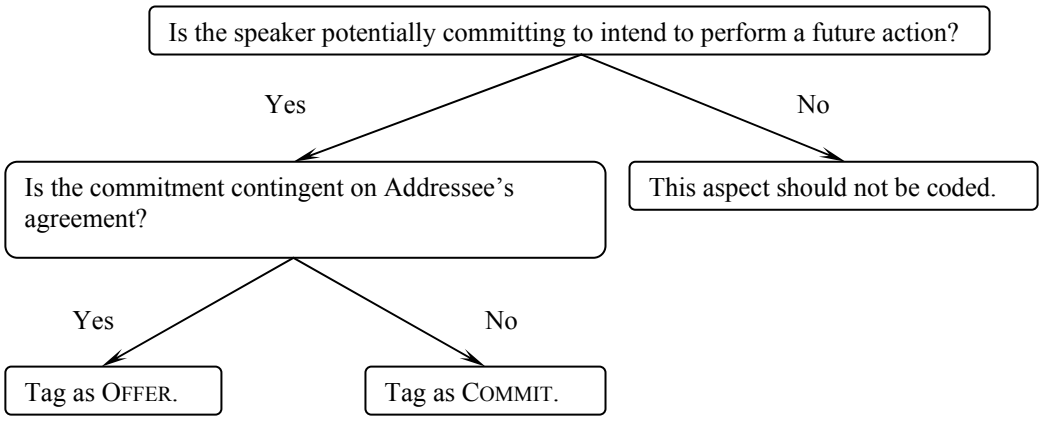


Figure 6.12 DAMSL decision tree for COMMITTING-SPEAKER-FUTURE-ACTION.

Backward-Looking-Function

AGREEMENT	a speaker's response to a previous proposal
ACCEPT	accepting the proposal
ACCEPT-PART	accepting some part of the proposal
MAYBE	neither accepting nor rejecting the proposal
REJECT-PART	rejecting some part of the proposal
REJECT	rejecting the proposal
HOLD	putting off response, usually via sub-dialogue
UNDERSTANDING	whether speaker understood previous utterance
SIGNAL-NON-UNDERSTANDING	speaker did not understand
SIGNAL-UNDERSTANDING	speaker did understand
ACKNOWLEDGE	demonstrated via back-channel or assessment
REPEAT-REPHRASE	demonstrated via repetition or reformulation
COMPLETION	demonstrated via collaborative completion
ANSWER	answering a question
INFORMATION-RELATION	how the content relates to previous content

An utterance has a backward-looking function if it refers back to a previous utterance in a dialogue. The different types of tags for this category of dialogue act are given above.

The decision tree for AGREEMENT is shown in Figure 6.13.

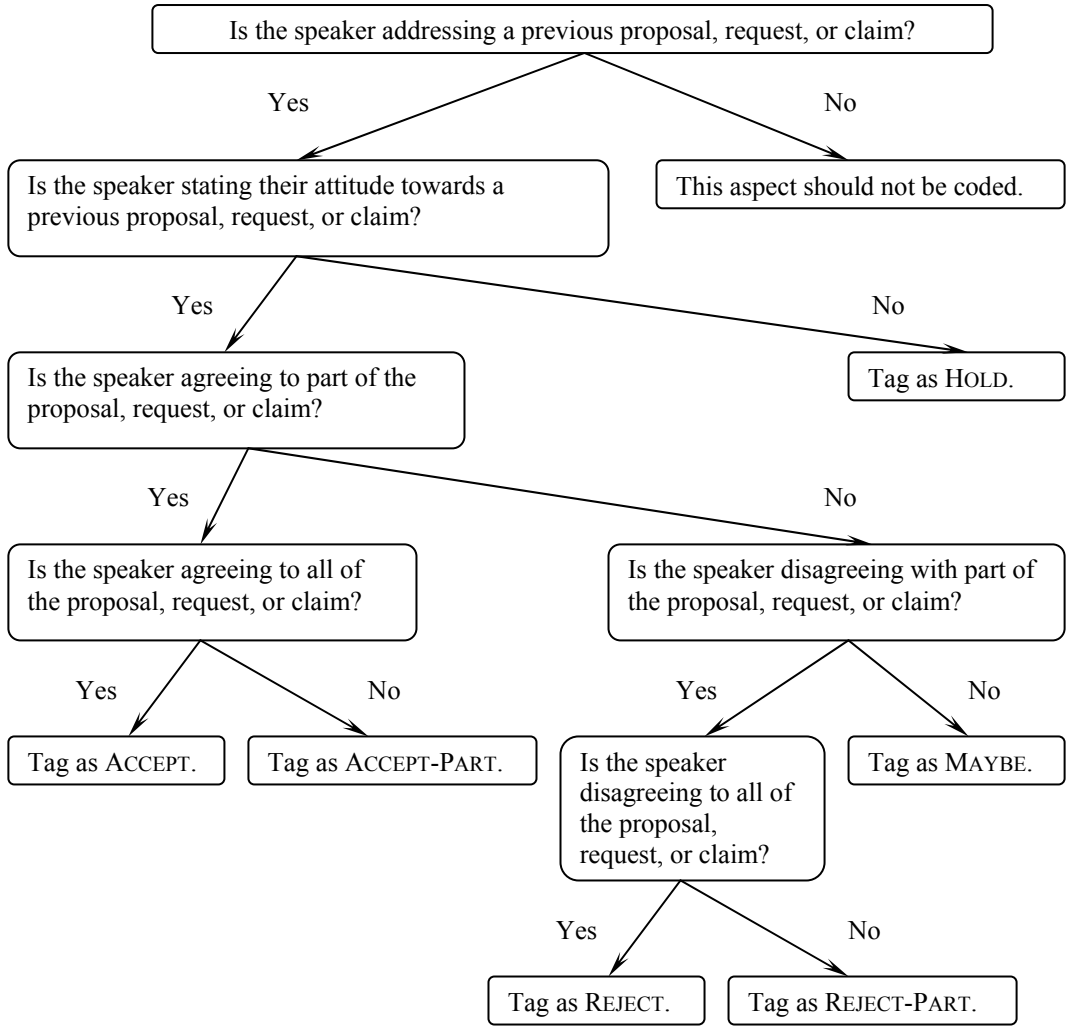


Figure 6.13 DAMSL decision tree for AGREEMENT.

This is a very quick overview of the DAMSL scheme to make the reader familiar with the kind of dialogue acts that are covered by it. The DAMSL mark-up has been used to tag corpora, such as the TRAINS corpus for example, to extract statistical models of which acts are most likely to follow each other (e.g., an OFFER is most likely to be followed by an ACCEPT or REJECT). Statistical models are built using *n*-gram, or hidden Markov model techniques; these methods yield the likelihood of one dialogue act following another by looking at sequences (of length *n*) of dialogue acts, and keeping a 'scoreboard' tally of possible combinations. This scoreboard 'weights' the choice of dialogue act by pattern matching with prior dialogue act interpretations (up to a 'depth' of *n* previous dialogue acts) in the current interaction in order to find the most probable dialogue act being performed. This is useful for intention-based, task-oriented

systems, as the statistical model informs the computational recogniser in cases when the dialogue act that is being performed is ambiguous (Core and Allen 1997a, 1997b, Core 1998).

There is one particular adaptation of the DAMSL scheme that is of most interest to us here, and that is the tag scheme that was constructed specifically with the annotation of the SWITCHBOARD corpus in mind. It is of most interest because it is, as far as I know, the only attempt to apply this scheme to non-structured, non-task-oriented, non-domain-dependent, (semi-) naturally occurring conversation.

The SWITCHBOARD corpus is a collection of 2,400 six-minute telephone conversations between strangers, who were required to chat informally to each other about a number of set topics (I shall give a more detailed overview of this corpus in Section 7.2.1.1). The SWBD-DAMSL annotation scheme was developed in order to apply the DAMSL tag-set to the types of dialogue that occur in the SWITCHBOARD corpus.

Originally the SWBD-DAMSL scheme was designed to be multi-dimensional, with about 50 basic tags defined, which could each be combined with other aspects of communicative acts (e.g. whether the utterance was related to task- or communication-management).

About 1,155 of the conversations in the SWITCHBOARD corpus (198,000 utterances, 1.4 million words) were labelled with this scheme in a project that was run at the University of Colorado at Boulder (Jurafsky et al. 1997a). The annotators, who were eight linguistics graduate students, used approximately 220 of the many unique combinations of the base tags and diacritics (i.e. extra tags denoting other communicative aspects).

This first attempt to tag utterances was not particularly successful, for two main reasons. Firstly because of the enormous number of possible combinations of dialogue act features, inter-annotator agreement was very poor. Secondly about 130 of the labels used (of the 220 different types) occurred less than ten times in the data. This meant that, for the purposes of statistical modelling of dialogue sequences, there was not enough data per class of acts.

In order to counter these problems, the SWBD-DAMSL scheme was ‘flattened’, so that the multi-dimensional diacritics used in the scheme were either subsumed by the base tags, or became dialogue annotation tags in their own right. Some of the base tags themselves were also grouped together under more general headings. The result was a much coarser-grained tag-set of about 42 ‘mutually exclusive’ utterance types (Jurafsky et al. 1997b and Stolcke et al. 2000). Table 6.2 shows the final list of dialogue acts identified, with examples from the SWITCHBOARD corpus and relative frequencies (Stolcke et al. 2000). Note that even after the process of flattening the scheme, the dialogue act distribution is still very much skewed. This is at least partly because the dominant dialogue act classes were not further sub-divided into task-independent, reliable criteria. Stolcke et al. (2000) claim that this is because there is always a

tension between the richness of expressibility, and the usability of a dialogue act scheme; in the SWBD-DAMSL scheme, usability is of prime importance.

SWBD-DAMSL Tag	Example	%
STATEMENT	*Me, I'm in the legal department.*	36%
BACKCHANNEL/ACKNOWLEDGE	*Uh-huh.*	19%
OPINION	*I think it's great.*	13%
ABANDONED OR TURN-EXIT	*So, -*	6%
AGREEMENT/ACCEPT	*That's exactly it.*	5%
APPRECIATION	*I can imagine.*	2%
YES-NO-QUESTION	*Do you have to have any special training?*	2%
NON-VERBAL	*[Laughter], [Throat_clearing]*	2%
YES ANSWERS	*Yes.*	1%
CONVENTIONAL-CLOSING	*Well, it's been nice talking to you.*	1%
WH-QUESTION	*Well, how old are you?*	1%
NO ANSWERS	*No.*	1%
RESPONSE ACKNOWLEDGEMENT	*Oh, okay.*	1%
HEDGE	*I don't know if I'm making any sense or not.*	1%
DECLARATIVE YES-NO-QUESTION	*So you can afford to get a house?*	1%
OTHER	*Well give me a break, you know.*	1%
BACKCHANNEL-QUESTION	*Is that right?*	1%
QUOTATION	*You can't be pregnant and have cats.*	.5%
SUMMARIZE/REFORMULATE	*Oh, you mean you switched schools for the kids.*	.5%
AFFIRMATIVE NON-YES ANSWERS	*It is.*	.4%
ACTION-DIRECTIVE	*Why don't you go first.*	.4%
COLLABORATIVE COMPLETION	*Who aren't contributing.*	.4%
REPEAT-PHRASE	*Oh, fajitas.*	.3%
OPEN-QUESTION	*How about you?*	.3%
RHETORICAL-QUESTIONS	*Who would steal a newspaper?*	.2%
HOLD BEFORE ANSWER/AGREEMENT	*I'm drawing a blank.*	.3%
REJECT	*Well, no.*	.2%
NEGATIVE NON-NO ANSWERS	*Uh, not a whole lot.*	.1%
SIGNAL-NON-UNDERSTANDING	*Excuse me?*	.1%
OTHER ANSWERS	*I don't know.*	.1%
CONVENTIONAL-OPENING	*How are you?*	.1%
OR-CLAUSE	*or is it more of a company?*	.1%
DISPREFERRED ANSWERS	*Well, not so much that.*	.1%
3RD-PARTY-TALK	*My goodness, Diane, get down from there.*	.1%
OFFERS, OPTIONS & COMMITS	*I'll have to check that out.*	.1%
SELF-TALK	*What's the word I'm looking for?*	.1%
DOWNPLAYER	*That's all right.*	.1%
MAYBE/ACCEPT-PART	*Something like that.*	<.1%
TAG-QUESTION	*Right?*	<.1%
DECLARATIVE WH-QUESTION	*You are what kind of buff?*	<.1%
APOLOGY	*I'm sorry.*	<.1%
THANKING	*Hey thanks a lot.*	<.1%

Table 6.2 SWBD-DAMSL tags and examples in frequential order.

The motivation for developing the SWBD-DAMSL tagging scheme was in order to extract a statistical model for the automatic recognition of dialogue acts in unrestricted conversation; such a model is based on the work of researchers in statistical modelling such as: Nagata and Morimoto (1994), Woszczyna and Waibel (1994), Suhm and Waibel (1994), Kita et al. (1996), Mast el al. (1996), Reithinger et al. (1996), Reithinger and Klesen (1997), Wang and Waibel (1997), Fukada et al. (1998) and Taylor et al. (1998). One of the incentives for recognising dialogue acts is to see whether knowledge of the dialogue act being performed can reduce the word recognition error rate in natural speech processing (i.e. of the raw acoustic signal).

Dialogue acts are recognised by means of three different knowledge sources: from pre-identified sequences of words (or linguistic cues), from prosodic contours that typify the performance of a particular dialogue act, and from a statistical model of the type of dialogue acts that are most likely to follow each other in spoken interaction (a variety of statistical dialogue act grammar in other words).

The dialogue act detection rate when using a combination of these models on automatically recognised words (the raw acoustic data) is 65% accuracy. This figure can be raised to 71% accuracy when the words are pre-transcribed and fed to the dialogue act detector in their corrected form. This, it is claimed, is not a bad labelling accuracy result when compared to a chance baseline accuracy of 36% (for example, if one just simply tags all utterances as STATEMENT, which is the highest occurring dialogue act in the SWITCHBOARD corpus data – see Table 6.2), and a human recognition accuracy of 84%.

This last statistic of 84% human accuracy in identifying dialogue acts is actually the inter-annotator agreement figure (which translates to a respectable κ statistic of 0.8). However, it is unclear to me that these numbers are comparable; the system annotator's dialogue act recognition capabilities are, presumably, calculated against the human annotators' standard, whereas the human annotators' capabilities are assessed by comparison to each other (rather than to some agreed standard). There are another couple of points that should be raised about this agreement statistic. Firstly, one could argue that it is still quite a low figure considering that the annotators are specialists in linguistics. Secondly, this is especially true since the annotators were allowed to confer freely during the annotation process[13]. How can it be claimed that a statistic on inter-annotator agreement is accurate, if annotators are encouraged to confer during the tagging process? The statistic cannot possibly be reliable as a measure of inter-annotator recognition of dialogue acts using the SWBD-DAMSL scheme.

---

[13] This was established through correspondence with Dan Jurafsky, who organised the annotation at the University of Colorado.

A factor that might have affected the agreement statistic adversely is that the annotators were not given access to the sound files of the various conversations they were expected to mark up with 'correctly' identified dialogue acts. This was presumably for reasons of speed (I suppose that it takes much less time to read a transcription alone than it does to manage an audio version of the conversation as well: play, rewind, replay, pause, annotate, etc.). The decision not to provide annotators with acoustic information is justified by Stolcke et al. (2000) by the results of the re-labelling of a portion (unspecified quantity) of the data, but with the audio version of the conversation available to the annotators. Results from this assessment indicated that at most only 2% of annotations would have been labelled differently (although annotators showed a greater tendency to change their annotations for AGREEMENTS and BACKCHANNELS, which are easily confused – annotation changes for these dialogue acts were 10%).

However, it seems counter-intuitive to accept that an annotator can label dialogue acts correctly with all auditory cues missing. This may have profound effects, for example, on the correct ascription of dialogue acts that have their force carried in intonation, such as questions that are performed declaratively (which are coded as a separate dialogue act type in the SWBD-DAMSL scheme). I would conjecture in such cases the lack of acoustic information is highly likely to produce discrepancies between different annotators.

In short there appear to be a few factors that render the annotator agreement statistic less than reliable as a test for the consistent labelling of the SWBD-DAMSL dialogue act tag scheme. I shall submit that the recognition rates themselves are also suspect, because of problems with the annotation scheme itself (see Section 7.3.2).

Stolcke et al. (2000) conclude that, while dialogue act modelling might in principle improve word recognition in speech, this was **not** a useful approach to use on the SWITCHBOARD data. They suggest that better results might be obtained using task-orientated dialogue, but this has yet to be proved.

Some of the main drawbacks with the statistical approach to dialogue act recognition are that that it is difficult to account for features of dialogue act recognition that rely on the long distance dependencies, discontinuous data and nested dialogue act structures that exist in human interaction. The identification of a dialogue act in conversation often depends on a variety of contextual information. Utterances can quite commonly refer back to other utterances made much earlier on in the conversation, or even to something in a previous conversation altogether (perhaps even on a different day). It is hard to see therefore how an *n*-gram, hidden Markov model, or any other statistical approach (such as Bayesian networks approach, see Pulman 1996, Keizer 2001), could capture this type of reliance on what has gone before. In the case of an *n*-gram model (for example, Stolke et al. 2000), increasing the size of *n* is no solution either, because the larger the number of items under consideration in the model, the more data is

needed to get meaningful results, the more time is needed to find a match for the data, and the less the likelihood is of finding distinct recurring patterns. Stochastic approaches assume a certain amount of predictability in conversational behaviour. While there are strict rules governing how we manoeuvre in a conversation, I would argue that the manoeuvres themselves are not statistically predictable. So my conclusion is that statistical models are not good enough to provide a realistic model of human behaviour on their own. In real conversation we are quite capable of jumping out of the expected and still being understood.

6.2.6 Conclusions about Structure- and Cue-Based Analyses

I have grouped the approaches discussed in Section 6.2 under the heading of structure- and cue-based analyses, because rather than focussing on the underlying intentions of the speaker to infer the meaning of an utterance, models have assigned meaning according to hearer expectations, and by using knowledge of the general structure of a dialogue, with clues to interpretation gleaned from typical cue-words and -phrases.

I believe this is a very promising approach, generally speaking, mainly because it provides shortcuts for some inferencing processes, and indeed in some cases side-steps the need to analyse a speaker's intentions at all. The most robust task-based dialogue systems developed to date have in actual fact proved to be those that combine both approaches.

Having said that a structure-based approach gives a good basis for the recognition of dialogue acts, it must be admitted that the recognition rates reported in the literature are not very high. This seems to contradict my assertion that structure-based analyses are promising for dialogue act recognition, but I believe this is not a weakness of the approach per se, but shows more a lack of rigour in its application. There are at least four criticisms that can be levelled at the research that has been carried out so far:

(1) There has been no attempt to show that all the dialogue acts that are actually performed in discourse have been fully represented in any theory or system.

(2) Although there is some correspondence between the different models of dialogue acts developed, some are fine-grained, while others are coarse-grained, and most are task-based, while very few are intended to be generic.

(3) There is no overall agreement about the sorts of act that are counted as dialogue acts, neither is there any account of how dialogue acts compound together to form an acceptable interpretation for an utterance.

(4) No dialogue act scheme has been developed from first-principles, with general conversation in mind.

Although the models have been structure-based, the approaches themselves (it appears to me) have been at times less than structured. This may seem like harsh criticism on my part, but I believe that it is this lack of rigour that has in part hindered the development of this approach to its full potential, especially for the treatment of naturally occurring conversation. I do not here wish to detract in any way from the very real achievements of these other dialogue models. Arguably it is only results that matter; if one is developing a system with a specific purpose in mind then success must be measured against the extent to which the project's aims are met. The moot question is, would these systems have benefited from a more generalised approach? It was in part trying to articulate and answer these observations that provided the motivation and justification for the development of my own model. In Chapters 8 and 9 I will begin to address these issues by introducing a more comprehensive framework from which to develop a structured theory of speech acts.

6.3 Brief Discussion

In this chapter I have divided the computational models of speech act recognition into either (1) *plan-/intention-based*, or (2) *structure-/cue-based* approaches. Of course in many respects these divisions are quite artificial, because approaches are generally mixed and include more than just one aspect of dialogue act recognition. What appear to be differences turn out to be reformulations of other theories. The relationship between different paradigms is often strong, either because they have been developed from other theories in common, or with the same background or overall goal in mind.

What is clear from attempts to apply some of the techniques described in this chapter to naturally occurring spoken dialogue, is that the frameworks devised for task-oriented dialogues do **not** translate to conversation. Partly this may be the impossibility of representing in a computationally tractable manner the amount of knowledge, both of how the physical world works, and about the social and cultural structures and conventions that are instilled in a 'normal' human being. But partly it is also that nobody has set out to study the structures of general conversation for themselves. The only attempt to apply dialogue act recognition to (two-participant) conversation is by Jurafsky et al. (1997b); even in this research, the dialogue act scheme was not constructed specifically with general conversation in mind, but adapted from a scheme designed for application to task-oriented dialogue (I shall be casting a critical eye over the SWBD-DAMSL scheme Chapter 7).

So, in short, the approaches available currently in the computational linguistic community are unsatisfactory for studying the structures of general conversation for a number of reasons:

(1) Nearly all the approaches (with the exceptions of DAMSL and CHAT[14]) are task-based at the application level.

(2) Nearly all mix up speech acts with moves that are rightly dealt with at other levels of abstraction, such as dialogue control, or social niceties (with the exception of DIT).

(3) All fail to provide good criteria for inter-annotator agreement; as a result, even those designed to be generic could not be consistently applied to natural conversation.

In the last chapters, I shall be tackling the questions:

(1) Why is it important to account for general conversation at all?
(2) What are the obstacles to studying general conversation?
(3) Why use speech act theory to analyse discourse?
(4) How can we start to come up with a general theory of speech acts?

---

[14] CHAT is a transcription system developed for use on the CHILDES (CHIld Language Data Exchange System) database of transcript data. The corpus mainly represents children's spoken dialogue for use in the study of language acquisition in children, through recorded conversations of children at play and through interviews with children of different ages.

Chapter 7

Speech Collection, Transcription and Annotation

> Perhaps the greatest single event in the history of linguistics was the invention of the tape recorder, which for the first time has captured natural conversation and made it accessible to systematic study. (Halliday 1994: xxiii)

Until fairly recently, the study of language was focussed by and large on its written form. Researchers did so principally out of necessity; before the advent (and universal availability) of devices for recording sound, there was no way of reproducing speech and conversation for study as it is really used by people in everyday life. Linguists and philosophers who did take an interest in spoken language were compelled either to rely on their own intuitions about talk, or to try and jot down conversations as they happened. Many chose instead to study the structures of language in their most readily available and easily examinable form: in written texts.

Arguably these days there is little reason not to, at the very least, refer to the real data to be found in recordings of spoken interaction[1]. In fact, some might go so far as to say that, with the dawn of video media, audio recordings are now not the most up to date means of studying spoken language either. (I shall be discussing this point further in Chapter 10.)

The question that is addressed first of all in this chapter is: why is it desirable to study spoken language interaction (specifically that which is socially motivated) at all?

7.1 Why Study *General* Conversation?

Before answering why it might be important to study general conversation, it would perhaps be worthwhile classifying exactly what I mean by *general* conversation.

7.1.1 What Counts as *General* Conversation?

Eggins and Slade give the following definition:

> We will define casual conversation functionally and, initially at least, negatively, as talk which is not motivated by any clear pragmatic purpose. (Eggins and Slade 1997: 19)

---

[1] I shall be discussing the various problems attendant on doing so in Section 7.2 when I look at the speech data needed to study spoken language phenomena.

Distinguishing between talk that is and is not pragmatically motivated is not at all easy. It is not clear that the word 'pragmatic' is here being used in the same way as I would like to use it. Eggins and Slade mean to contrast between formal, transactional settings (where the main aim is to achieve some obvious goal) and informal, casual ones (where the main aim is socialisation). But arguably all conversation is motivated by pragmatic purpose at some level, even casual chitchat.

I shall be assuming that there are (at least) the following different types of settings in which communication can occur (see Table 7.1, as classified by Clark 1996: 8):

Type	**SPOKEN settings examples**
Personal	A converses face to face (or over a telephone) with B.
Non-personal	Professor A lectures to students in class B (monologue and power imbalance).
Institutional	Lawyer A interrogates witness B in court.
Prescriptive	Groom A makes ritual promise to bride B in front of witnesses.
Fictional	A performs a play for audience B.
Mediated	C simultaneously translates for B what A says to B.
Private	A talks to self about plans.

Table 7.1 Types of spoken settings[2].

It is the *Personal* category above that I wish to associate with general conversation, and no other. I have narrowed down my research to this category alone on the assumption that it subsumes all other categories. The decision to restrict the scope of communication type under consideration in this way is supported by a variety of arguments:

(1) For the majority of people, their main language use is face-to-face conversation; for those of the world's population who are illiterate, this is almost the only means of communication (with the exception of gestures and signals).

(2) Most of the world's languages have evolved in a spoken environment – the written form of language is a relative newcomer on the linguistic scene.

(3) Face-to-face conversation is also the means by which children first acquire language (as well as other communicative and social skills). If we acquire language by the method of face-to-face interaction, then it is not unreasonable to suppose that all other styles of communication have sprung from, and are a variety of, it.

---

[2] Clark includes a written version of the settings examples for each category, but this is omitted here, as I am not concerned with written communication in my research at all.

∴ There seems no point in only studying a form of language that does not provide an account of face-to-face conversation, which is what theorists in linguistics have done for many years; as I have already mentioned in the introduction to this chapter, the trend for looking at spoken language as it is actually used, is a relatively recent one. Even though formerly scholars may have claimed to be analysing actual utterances, these were generally composed of made up examples produced to illustrate a particular theoretical point rather than taken from real conversation. The theoretical points were thus based on a theoretician's intuitions rather than from observed behaviour in spoken data. This practice has continued to date – particularly in the field of generative linguistics – even after the advent of easily available tape recordings in the 1950s (and these days, video recordings)[3].

Equally, developing a partial theory of communication patterns, that is only applicable to a task- and transaction-based subset of a more generic model of language use, also seems highly unsatisfactory.

So if face-to-face interaction is key to language evolution, then it can be argued to be the basic setting from which other uses of language are derived. It is this last point that led me to study conversation in its least rigid setting; because, I believe that behaviours in other settings are simply modified from the base case. In other words, conversation represents the most generic form of language, which is then constrained in different ways to fit the circumstances of the current talk exchange, whatever that may be; it is the foundation from which our other language use extends. Therefore, if I can find a generalised model of speech acts at this level, it should be possible to then apply it to other situations by means of a set of behavioural heuristics.

In the following discussion, I shall characterise general conversation as conversation that is inspired and carried out at least principally for the personal and social gratification of the participants. As nearly as possible, this should be between two equals, and should be contrastively different to the genre of transactional conversational behaviour (such as booking tickets at a box office, asking for timetable information, opening a bank account, etc.).

7.1.2 Language as a Social Medium

One can argue specifically for the study of general conversation for a variety of reasons as I have done in the previous section. I have chosen to focus on the structures of spoken dialogue mainly because it is a genre that is largely overlooked and under-represented in modern day

---

[3] I must admit that this is hard to avoid at times. I have myself in this very dissertation resorted to using 'made up' examples when I have been unable to find an instance of the kind of behaviour that I wish to exemplify in the collected corpus of conversations. I have tried wherever possible to use real data however.

research in speech act theory. The study of casual conversation is usually associated with the work of socio-linguists, psycholinguists, studies of child language acquisition, dialectology, etc. but is rarely thoroughly represented in computational linguistics[4]. Possibly this is with good reason, as conversation covers an enormous range of phenomena, and is often motivated and produced from purely social impulses, which are difficult to model computationally. With this in mind, it is perhaps unsurprising that computational linguists have chosen to cut down the field of study in some way and to some degree: in order to make the process of automation more manageable. Partly no doubt this is due to the many difficulties presented simply in collecting the data to be studied in the first place (which I shall be describing in Section 7.2). But also this is because it is the transactional type of talk exchange that it would be of most immediate use to be able to automate.

The study of general conversation has tended to concentrate on various salient features of this genre, such as turn-taking behaviour, or the use of certain kinds of discourse units, without looking at how the patterns of different levels of language (word, clause, utterance, turn) interact with each other to produce the overall meaning of general conversation. Not only is an analysis of the micro-interactions missing from current research, but also how these play a part in the development of our conversational structures at the macro-level.

Fundamentally conversation is used to carry out social activities; language as we know it would not exist if it were not for the social activities it motivates. Understanding how conversation works is important because we use it to position ourselves within a society; we manipulate our language in order to be able to do that. We use casual conversation to align ourselves with speakers (often against, or at the expense of, some other person), to gain social 'power', or confirm one's role in a family group. We use language as a vehicle for displaying and carrying out certain social behaviours. Although the social role of communication is at a level removed from my interests here, it is still of relevance; explaining how our linguistic skills allow us to

---

[4] A notable exception to this is the increasing interest in building personable 'chatbots', computational agents that will converse with users either through the medium of text or through speech recognition and synthesis. Microsoft, for example, has been working for some time on such agents to provide help for using their software (Ball et al. 1997). One example of a chat program is CONVERSE (Batacharia et al., in Wilks 1999), which was the Loebner prize-winner in 1997, and attempted to fool users into believing that they were chatting textually with another person. CONVERSE realised this through the strategy of retaining as much control of the flow of conversation as possible, so that the user would not have the chance to talk about topics that the program did not have specified in its predefined scripts. CONVERSE is amongst a host of chat programs built in the tradition of ELIZA (Weizenbaum 1976) and PARRY (Colby 1973). The most recent Loebner prize-winner (three times winner in 2000, 2001 and 2004) is the chatbot ALICE, written in XML-based AIML (Artificial Intelligence Mark up Language), which uses an extensive semantic network. None of these chatbots is convincing over a significant period of time.

display these behaviours is an important line of enquiry from the point of view of explicating patterns of social strategy.

It is mainly in our social communications that we build our beliefs and frame our mental image of the world. It is in speaking that we shape our inner perception of reality, whether we are aware of it or not. In a sense we might say that language itself is the mind's librarian and archivist. Berger and Luckmann (1966: 172-3) have the following to say about this position:

> The most important vehicle of reality-maintenance is conversation. One may view the individual's everyday life in terms of the working away of a conversational apparatus that ongoingly maintains, modifies and reconstructs his subjective reality… It is important to stress, however, that the greater part of reality-maintenance in conversation does not in so many words define the nature of the world. Rather, it takes place against the background of a world that is silently taken for granted.

I would argue that for this reason alone, the study of informal conversation is crucial. Our beliefs and opinions are shaped by, and given expression through, our daily interaction with other people. To ignore this means of communication is to discount the greatest informational and social resource in our every day life. It is because of the importance of general talk to human development that it is so surprising that research into its structures is so grossly underplayed and under-explored.

7.1.3 Features of Face-to-Face Spoken Interaction

Because speech is immediate both in production and understanding, it has many features that are unlike those to be found in written texts. In Figure 7.1, I show twelve different qualities that epitomise face-to-face conversation (I have adapted and added to this list from Clark 1996: 9).

To this list one might also add: (specific to non-transactional conversation) *Continuity*, *Informality*, (general to all spoken interaction) *Context Dependency* and *Feedback Expectation*. I shall take a look at some of these features in more detail in the following discussion. I do not necessarily conform to the order in Figure 7.1, but try to show how the different features are interrelated with, and in some cases dependent on, each other.

Immediacy	1)	*Co-presence*	The participants share the same physical environment.
	2)	*Visibility*	The participants can see each other.
	3)	*Audibility*	The participants can hear each other.
	4)	*Instantaneity*	The participants perceive each other's actions at no perceptible delay.
Medium	5)	*Evanescence*	The medium is evanescent – it fades quickly.
	6)	*Recordlessness*	The participants' actions leave no record or artefact.
	7)	*Simultaneity*	The participants can produce and receive at once and simultaneously.
	8)	*Contractedness*	The message is often not given in full but left to the hearer to interpret.
Control	9)	*Extemporaneity*	The participants formulate and execute their actions extemporaneously, in real time.
	10)	*Dynamism*	Control of the interaction is negotiated while it is taking place.
	11)	*Self-determination*	The participants determine for themselves what actions to take when.
	12)	*Self- expression*	The participants take action as themselves.

Figure 7.1 Features of face-to-face conversation.

Co-presence & Visibility: Unlike in written language, in speech participants can refer to the everyday objects that surround them in their conversation. The physical and visual context is very important for understanding an act of spoken communication. In fact, some conversations cannot be interpreted fully without knowledge of the visual cues contingent on the utterances themselves. Take for example the case of someone pulling a face when asked to carry out some task (the washing up say). This then provokes a response from the requester, without the requestee having uttered a word. Sometimes the participants might be talking about some other visual stimulus within their current physical context, depriving a person who is not actually present in the same place from a complete understanding of the conversation. Of course human beings have an excellent facility for imagination, and can furnish for themselves a hypothesised context in order to impose some kind of reasonable (although not always correct, or even adequate) interpretation for what they are hearing. This is what I did in Chapter 1 when I hypothesised a context for Albert and June from my limited knowledge of the situation of the conversation. This shows the highly ***context dependent*** nature of spoken language. Context dependency poses severe difficulties for the thorough analysis of recorded and transcribed

conversation, as it is especially vulnerable to loss of context. People go in and out of rooms, refer to objects visible at the time (but not visible to the analyst), etc.

It is not only the world surrounding conversationalists, and their facial expressions, that will add meaning to the interpretation of their utterances, but often also their gestures. Speakers send many visual signals that accompany their speech to transmit rhythm and emphasis to phrases and words respectively. Sometimes gestures will replace utterances or parts of utterances altogether. So speakers not only receive and produce sound messages ***simultaneously***, but they also manage to incorporate the visual message as well. It is for these reasons that I will be arguing, after the fact, that perhaps the most appropriate method for studying speech without recourse to the immediate physical environment, is to analyse telephone conversation, in which the speaker and hearer themselves are stripped of their visual input. There are, however, problems with recording telephone conversation relating to issues of protecting a person's right to privacy.

Evanescence & Recordlessness: Speech is not subject to inspection or perusal by the receiver. Hearers in a conversation rely on their often less than perfect memory to recall what other people have said. An interesting question is: how do people store this information? However it is done, the human facility for recovering information is quite efficient; we are more than capable of rephrasing or reciting the gist of a conversation, even at some distance in time from when the conversation took place. Quite often the fine-grain of conversation is lost, but the overall structure is retained. The way that people reproduce the gist of a conversation, and our ability to summarise 'in other words', may provide important clues about the form in which we store our information. One of the strong points of the approach I have taken (to be described in Chapters 8 and 9) is that it provides an explanation of how the important information might be retained from a dialogue in a compact form.

Audibility & Feedback Expectation: Speech is not only subject to decay (of information over time due to faulty memory), but also subject to being misheard or misunderstood in the first place. This is one of the arguments against the idea of a mutual contextual belief, as there is no guarantee that the message has reached the hearer intact. People often rely on feedback from hearers to check that the message has got across. It is a hearer's responsibility to demonstrate understanding of the topic of discussion.

Signalling understanding is important in maintaining the ***continuity*** of a conversation. Often the hearer does not need to take over a whole turn in order to reassure the speaker that he has understood (regardless of whether this is actually the case or not). Continuing contributions, when the hearer does not want to take over the turn, but wishes to indicate their continuing acceptance of, and attention to, what the speaker is saying, have the following characteristics:

(1) *Acknowledgements*: 'I see', 'm', 'gosh', 'really' are all minimal so that they do not interfere with the flow of talk from the speaker (and all more or less stand for 'Yes, I understand what you're saying').

(2) *Scope*: The hearer marks what they are accepting by placing the signal at or near the end of the section of speech (or utterances) with which they are agreeing.

(3) *No turns*: The hearer does not need a proper turn in order to do this.

(4) *Overlapping*: Commonly exemplified by the acknowledgement coinciding with the end of the acknowledged utterance.

(5) *Backgrounding*: Acknowledgements are brief, quiet and simple. In speech, 'm', 'uh huh', 'yes', and 'yeah'; in gestures nods and smiles.

These continuing contributions show the ***simultaneity*** and co-operativity of spoken conversation. Other signals of understanding are:

(1) *Unison completion*: (Tannen 1989: 60) This is a variation of backgrounding, when the hearer joins in with the last words of the speaker's utterance to show they have understood so well that they can even finish off the utterance using the same words as the speaker uses.

(2) *Collaborative completion*: (Goodwin 1986, 1987, Lerner 1987) When someone pauses, to look for the right words perhaps, and their hearer finishes off a sentence for them. The completion is then accepted or rejected. In the former case the conversation would then pick up where it left off with the speaker regaining their turn and resuming what they were saying; in the latter case the speaker would finish their utterance in the way they actually intended to.

(3) *Truncations*: When the hearer interrupts as soon as they understand what the speaker intends in the interests of curtailing an interaction (sometimes the speaker even invites such an interruption).

(4) *Repetition*: Sometimes a simple repetition of the information, or a salient part of the whole message that has just been received will be sufficient to reassure the speaker and the hearer that the hearer has heard the utterance correctly.

(5) *Summary*: Rather than just repeating the message word-for-word, a hearer might have a preference for rewording the message and repeating a paraphrased version to convince the speaker of his understanding.

(6) *Adding extra information*: A hearer also has the option of expanding the content of the utterance, which is not only a good indication of understanding, but also a method of adding to what is said in some way.

These are all strategies for conveying one's understanding and speeding up the process of communication. Hearing what is said in the first place is obviously an essential prerequisite for understanding, which is why most utterances require some kind of reaction from the hearer. If a hearer fails to respond appropriately, it is often difficult to determine why. Brown (1995: 34)[5] gives the following list of reasons:

- The listener was not listening to what was said (or did not hear what was said).
- The listener heard what was said, but was so engrossed in interpreting a previous utterance that no immediate further processing of the current utterance was possible.
- The listener heard what was said but did not understand what the utterance (or some part of the utterance) meant, for instance did not know the meaning of one of the words.
- The listener understood the words of the utterance and parsed it correctly but could not interpret 'the thin meaning' in the current context and was waiting for more information before trying to respond.
- The listener understood the utterance in the current context of information but was unable or unwilling to produce an appropriate response.

So we see that these features of spoken, face-to-face interaction are pivotal to understanding, not only the talk itself, but also the structure of conversation.

Contractedness: One of the main problems with studying spoken language (recorded and transcribed) is that, unlike written language, much of the information is condensed and re-used. Written language for the most part is by nature self-contained. Of course written texts do make use of internal references as well (co-reference, anaphora, ellipsis, references to external entities, beliefs, etc.), but not in the same way as spoken language does. One has only to consider the use of the word 'Yes' to realise that spoken interaction is very different to that of written prose. 'Yes' can almost be said to have no meaning at all outside of a spoken context. Because of the ***extemporaneous*** and ***dynamic*** nature of spoken language, the use of linguistic shortcuts to avoid repetition of the same piece of information, or the contraction of the answer to an indirect request (e.g. 'Can you tell me the time please?', 'Ten past two') is extremely

---

[5] In Chapter 8 I shall be looking at a more principled explanation of how errors can occur in communication, and how we use our knowledge of the way speech acts work in order to fix them.

useful. However, they cause problems for the conversational analyst, and sometimes also for the hearer himself, when the reference is ambiguous.

Informality: Spoken language is less formal and much more likely to be ungrammatical than written language. Corrections are often made on the fly and are so much a part of our processing skills, that often they go to all intents and purposes completely unperceived. Utterances are often started, left unfinished, restarted and rephrased because the planning is going on at the same time as the production. Sometimes mistakes are made and left uncorrected; it is left to the hearer to infer what the speaker actually meant to say. I have some anecdotal evidence of these phenomena. In the process of collecting some of my own recordings, one of my subjects was deeply shocked on reading the transcription afterwards. He said that he had not realised how incoherent and broken his speech was. It is unlikely that it was perceived as so by his audience however who were able to track the changing output as it occurred. This demonstrates the ***instantaneity*** of speech production and understanding.

Having discussed what counts as general conversation, why it may be important to study its structures and what are its distinguishing features, I now turn to the problem of how it is possible to study a form of language that is inherently so evanescent in nature.

7.2 Speech Data Sources

For one trying to investigate features of spoken language, there are really only two paths to follow; one must either use an available spoken dialogue corpus, or collect and transcribe one's own data. The former option is obviously preferable for time- and labour-saving reasons. However, there are many obstacles to be overcome if one is to use speech corpora for research into spoken language phenomena. In this next section, I will consider the problems attendant on these various options, and indicate at each stage the grounds for my choices. I shall first discuss the available spoken language corpora, then the collection of my own data for transcription.

7.2.1 Speech Corpora[6]

At the start of this research, I had hoped to use existing spoken dialogue corpora for my investigations into speech act use and dependencies. However, this proved to be less easy than I thought. To begin with, the numbers of freely (or at least reasonably freely) available spoken language corpora are relatively few. I detail here the main relevant (English) corpora that were found, and the reasons why they did not fulfil my requirements.

---

[6] All information given here is up-to-date up until 2003.

7.2.1.1 Linguistic Data Consortium

The United States Defence Department's Advanced Research Projects Agency (DARPA) made a policy change in 1986 in an effort to centralise resources for its speech research programs. The success of this data sharing led to rapid progress in the areas of speech recognition, message understanding, document retrieval, and machine translation. The Linguistic Data Consortium (LDC) was founded on the back of this success in 1992, and has provided a useful forum and resource for large-scale development and widespread sharing of data for both academic and industrial researchers in emerging linguistic technologies. There are currently four spoken language corpora available from the LDC that are directly relevant to my research:

(1) CALLFRIEND: The corpus is made up of 60 unscripted telephone conversations of between 5 and 30 minutes duration. Information is included about the participants (sex, age, education, etc.) and the individual call (channel quality, number of speakers, etc.).

(2) CALLHOME: The corpus consists of 120 unscripted telephone conversations of up to 30 minutes duration. Out of the 120, 90 calls were made from North America to family and friends overseas; 30 were placed from and to a North American location.

(3) SANTA BARBARA CORPUS OF SPOKEN AMERICAN ENGLISH (SBC): The corpus is based on hundreds of recordings of naturally occurring speech. Part 1 (the only release so far) is made up of 14 dialogues of between 15 and 30 minutes duration. The recordings were made in a wide variety of locations all over the United States, representing people from different regional origins, ages, occupations, and ethnic and social backgrounds. The contents of the conversations themselves are a realistic reflection of the way people use language in their everyday lives and comprise: general conversation, chitchat/gossip, arguments, disagreements, job-related talk, financial meetings, a classroom lecture, a political conversation, a mathematics tutorial. The privacy of the individuals recorded has been respected by the removal of the original personal names, place names, phone numbers, etc. from the transcriptions, and the filtering of the audio files to make the same information irrecoverable. This was the most promising corpus of all those obtainable from the LDC (price-wise especially), but proved problematic to use for two reasons: (1) because conversations are taken out of context (both in terms of personal/social context, as well as cultural), and (2) because there was no restriction on the number of participants in the conversation, the instances of unclear conversation were often quite high.

(4) SWITCHBOARD: The SWITCHBOARD corpus consists of about 2400 telephone conversations of 6-minute duration. There were 543 different speakers (302 male and 241 female) from all over the United States, who were strangers to each other. The callers would dial into an automated telephone exchange, and a computer-driven operator system would prompt the

caller for recorded information (such as selecting a topic about which the participants would speak) and select an appropriate 'callee'. The speakers could then talk to each other until they finished their conversation about the topic and ended the call. There were about 70 different topics from which to choose; the only constraints on the participants were that they would only talk to the same person, and use the same topic, once.

7.2.1.2 The Survey of English Usage

Lord Randolph Quirk originally founded The Survey of English Usage in 1959. The first corpus consisted of one million words of written and spoken British English produced between about 1955 and 1985. The corpus was first released in the form of grammatically annotated slips of paper, but then later computerised. In 1983 Sydney Greenbaum took over work on the Survey, which was responsible for the co-ordination of the International Corpus of English (ICE) project, and also for the British component (ICE-GB) of the same. The aim of the ICE project was to collect as many varieties of English from around the world as possible for comparative purposes. The Survey produced the grammatical and syntactic annotation schemes, as well as a number of software packages to aid in the production of the ICE corpora. ICE-GB boasts that it contains the largest quantity of parsed spoken data in the world.

Spoken Texts (300)	Dialogues (180)	**Private (100)**	face-to-face conversations (90) phone calls (10)
		Public (80)	classroom lessons (20) broadcast discussions (20) broadcast interviews (10) parliamentary debates (10) legal cross-examinations (10) business transactions (10)
	Monologues (100)	**Unscripted (70)**	spontaneous commentaries (20) unscripted speeches (30) demonstrations (10) legal presentations (10)
		Scripted (30)	broadcast talks (20) non-broadcast speeches (10)
	Mixed (20)		broadcast news (20)

Table 7.2 Spoken texts in the ICE-GB corpus.

INTERNATIONAL CORPUS OF (BRITISH) ENGLISH (ICE-GB): The spoken component of ICE-GB consists of a total of 300 'texts' of about 2000 words each, which date from between 1990-6. A break down of these is shown in Table 7.2. We can see that of the 300 spoken texts available, it is only a third of these that would be of direct significance to my research.

7.2.1.3 The British National Corpus

This project to gather a very large corpus of a wide range of modern British English was managed by a consortium of industrial and academic institutions, led by the Oxford University Press. The aim was to provide a resource to cover as broad a cross-section of British society as possible, which is marked up using internationally recognised standards of text encoding, so that the generality of the corpus would make it accessible for many different research purposes. Data was collected between 1991 and 1994, and the BNC was first disseminated in 1995.

> BRITISH NATIONAL CORPUS (BNC), including the CORPUS OF LONDON TEENAGER ENGLISH (COLT): The corpus comprises 4,124 texts of which 863 are transcribed from spoken conversations or monologues. The spoken component of the BNC accounts for about 10% of the total corpus. As there are just over a hundred million words in total, the spoken word count is around ten million.
>
> The spoken component of the BNC is further divided equally into two different types of speech data: a **demographic section** that contains recordings of natural spontaneous conversations made by members of the general public, and a **context-governed** section that includes speech produced under more formal settings (such as classroom interactions, business meetings and interviews, sermons, speeches, sports commentaries, radio phone-ins, etc.).
>
> It is the demographic content of the BNC that is of interest for my research. Volunteers from the general public (about 124, from 38 different parts of the country) made recordings on personal stereos. Their numbers were an even distribution of various factors: different social classes, gender and age.

7.2.1.4 Evaluating the Existing English Speech Corpora

In the previous sections of 7.2.1, I have surveyed the English speech corpora available, and described the features of each and the methods used for their collection. In this section, I shall compare the different corpora, and explain why it is that most are practicably unsuited (for one reason or another) for analysing speech acts as they occur in spoken dialogue.

A comparison of some of the important features of the six corpora that are described above, is shown in Table 7.3.

Corpus Name	Type of English	Speakers	Price	Data Source	Audio?
CALLFRIEND	U.S.	2	$600	Telephone	✓
CALLHOME	U.S.	2	$1000[7]	Telephone	✓
SBC	U.S.	2+	$75	Microphone	✓
SWITCHBOARD	U.S.	2	$2000	Telephone	✓
ICE-GB	U.K.	1+	£294–586[8]	Various	→[9]
BNC	U.K.	1+	£250	Various	×

Table 7.3 Summary of features of spoken corpora.

One can make some general observations about the characteristics of these corpora. The first is that none are available free of charge. This reflects the fact that such corpora are extremely expensive to collect in terms of man-hours; huge quantities of data are required to satisfy the demands of speech processing computer programs, in order to build robust lexicons and grammars. Not only is obtaining the data in the first place costly, but once it has been amassed, there are the additional costs of transcription, documentation, maintenance and distribution.

Arguably the studying of speech act recognition could most effectively be carried out, not only in one's native tongue, but also in one's own country's dialect (simply from the point of view of being able to understand the references). This would rule out the study of American English data for a British researcher (and vice versa); but it is only the U.S. corpora that come with the original audio data (which in most cases is also time-stamped and aligned to the transcription).

Setting aside the various considerations I have just discussed, the main problem with these corpora is that they were each designed with a particular application in mind. So, while they fulfil their intended purpose, they are not particularly suitable for a detailed study of conversational structure:

CALLFRIEND:	Language identification.
CALLHOME:	Speech recognition.
SBC:	Unspecified.
SWITCHBOARD:	Speaker identification, speech recognition.
ICE:	Comparative studies of varieties of English in the world.
BNC:	Lexicography, literary studies, and corpus-based linguistics.

---

7 For the audio files only; the transcription has to be bought separately.

8 Depending upon the type of licence required.

9 The digitised audio files, aligned with the transcription will be available with the second release of ICE-GB.

Even with the availability of these resources, a large gap in the data can still be seen. There is, in my opinion, an urgent need for a sizeable corpus of naturally occurring (unscripted), private, face-to-face (or telephone) dialogue, which is transcribed and annotated at various levels and linked to digital audio files. Especially useful would be the collecting of more than just the background facts about the speakers themselves; explanations about the content of the conversations is also vital in order to have the least hope of fully understanding what is said in context. Such a corpus simply does not exist, and would take an enormous amount of man-hours to collect together. As this is far beyond the means of one person I have had to content myself with using some of the available corpora, and also taking a few recordings myself. I realised very quickly that I was able to comprehend and work much better with my own recordings than with the de-contextualised material collected by others. This posed a dilemma, as building my own mini-corpus was costly in terms of time and effort.

7.2.2 Collecting Recordings

The conversations that I recorded myself all took place between 1997-1999. I tried to choose situations that were as natural as possible to enable me to collect conversations that are genuinely and spontaneously produced. The participants involved were all aware that they were being recorded (a factor which would affect and constrain the content of the conversation in some instances).

Because I am interested in analysing the underlying intentions and functions of utterances in context, I felt it was important that I be present at, and participate in, the conversations that I collected. There are obvious methodological drawbacks to this set up:

(1) As I am both participant and analyst, I might unconsciously be affecting the course of the conversation, thereby also invalidating any conclusions I might draw from the study of such conversation. While this is a fair point, I do not think that the features of conversational construction that form the basis of my study can be controlled at a conscious level when speaking. People are so used to just talking to each other naturally that even the unnatural presence of a tape recorder is forgotten after a short while once the participants get absorbed in their conversation. I would argue that this is true just as much for myself in my role as a conversationalist as for the other participants.

    A greater risk is that in the analysis itself, I may add elements of meaning to the interpretation that were never intended in the case when I feel that my memory of the situation gives me a better knowledge than the recording on its own. There is no real defence against this criticism; the best I can do is to make sure that the model I develop is defined in a theoretically rigorous manner, and applied to the transcribed conversation consistently.

(2) As my interest is in the way that people *understand* utterances, I have chosen to record the conversations of my peers and my family. This is obviously an unbalanced group of people in terms of variation of socio-economic background, ethnicity and nationality. I justify my choice of studying the language of those around me from the standpoint of easy access to the participants, knowledge of their backgrounds and shared experience. I am more likely to be able to understand and analyse a participant's motivations for producing an utterance if I know about them (although see the caveat in (1) above). When investigating the possibility of using pre-existing corpora for my observations (as discussed in Section 7.2.1), I noticed that identifying the speech act being performed with any certainty was often impossible, precisely because it was difficult to know exactly what was going on in the situation. This was true even of the transcriptions that were accompanied by the audio files.

Part of the problem with available spoken corpora is that they have been collected with a certain research ethos in mind. Spoken corpora are mainly used by those wishing to exploit surface linguistic information, those who run stochastically informed (e.g. part of speech) taggers and parsers to extract widespread features of language, for a variety of purposes (note that I am not claiming here that this is the only possible use of the collected spoken language corpora, just the most common). This kind of quantitive feature extraction process is not of much use to my thesis that there is an underlying grammar of speech function that is determined in context.

For academic observers and researchers there is often a dilemma to be faced. The easiest way to collect conversation is to tape one's peers. But then one can only say that one has started to study and represent the psychology of what Brown (1995) calls 'me and my friends'. This harks back to Davidson's suggestion mentioned earlier (in the discussion in Chapter 2) that we are only capable of understanding language that uses similar concepts to our own. I accept that this is a criticism that can be levelled at my own research. There are good arguments for recording the interaction of strangers for analysis (as did projects such as MAPTASK and SWITCHBOARD). Friends will often be able to shortcut the communication process due to shared experiences and background, thus perhaps creating conversations that are massively under-specified and hard to follow for the outsider (this is the very reason why I found that analysing other corpora of conversations between friends extremely difficult). However, to study intention I would argue that some kind of self-analysis is necessary; surely the only person who can tell what I meant to say, or at least what I think I meant to say, is me. By studying the interaction of strangers one inserts an extra level of obscurity between the researcher and the analysis.

(3) Collecting and transcribing conversations is extraordinarily time-consuming and labour intensive. The transcription of a ten-minute conversation may take more than a whole day,

even when one was present at the conversation so has the benefit of all the contextual clues to meaning and interpretation. In consequence, the number of my own recordings of conversations that I have used to test my model is not considerable. I cannot claim therefore with any statistical certainty that the model I have postulated is a true reflection of what actually happens in conversation, only that it appears to match the data I have looked at from the comparisons I have made to date. My observations here can only be said to be qualitative; I do not have the quantitative results to back them up. It is up to someone else to take my work and apply it more widely to a greater range of conversational situations and to assess my conclusions quantitatively – I have not the resources to do so myself here.

At the beginning of my research, I looked at conversations that took place between myself and:

(a) Male friends (one at a time).
(b) Two female friends.
(c) Three other members of my family.
(d) More than ten other members of staff at academic meetings.

I was at first attempting to incorporate many different types of talk settings and different levels of formality. I decided that this was too wide a field, and that the study of general, informal conversation alone was sufficient a challenge, so I cut out option (d).

When transcribing conversations that occurred between more than three people, I noticed that this often caused a split in the conversation so that more than one strand would take place simultaneously and concurrently overlapping in time. This made the job of transcription nearly impossible, so that these conversations were also ignored (at least at the point where the split happened). So recordings were restricted to a maximum of three participant conversations, in order to be assured of the best possible clarity of speech for transcription.

This, incidentally, is one of the problems that I noticed while investigating other speech corpora. The greater the number of people present during the conversation, the more likely that utterances will be lost in transcription due to cross conversations, overlapping speech and interruptions. I discovered that, for the purpose of studying spoken language, two, or three participants at most, form the ideal group of interactants. From the point of view of modelling discourse, it is the three participant conversations that are of most use. This is because the combination of utterances that three people can produce is representative of all the possible interaction patterns that can occur between different speakers[10], but there are not enough people

---

[10] I hope to show exactly what I mean by this in Chapter 9.

for the conversation to split up into different groups (and thereby mar the potential for good clear recordings and transcriptions).

7.2.3 Transcription

The transcription process itself throws up many representational difficulties. There is significant loss of information in the process of transcribing a conversation into some kind of orthographic representation. We lose certain features such as voice quality, whether the speaker is interested, amused, bored, etc. (i.e. attitude of expression), volume (in a heated debate for instance), emotion, intonation (although this can be represented perhaps in a very crude way by punctuation), speed, pauses, hesitations, stutters, mispronunciations, as well as elision, assimilation, omission of sounds, or lengthening of syllables, etc. We also lose all visual and physical contextual clues, such as the speaker's facial expression, as I have already mentioned.

The process of transcription is by its very nature based on the speaker, which ignores what goes on with the listener (for example when signalling a wish to speak). But the act of listening may well be vitally linked to planning the hearer's next contribution as speaker. This is one of the reasons why multi-modal analysis is so much in vogue these days, so that researchers can take into account and have access to the visual context as well as just the audio reproduction. Perhaps, as I mentioned earlier, there is an argument for studying telephone conversations in the analysis of purely spoken interaction, as this medium naturally eschews visual signals.

It is the loss of information inherent in the process of transcribing spoken interactions that has motivated my insistence on the necessity of working (at the very minimum) with the original recordings, to at least partially inform the further analysis and annotation of the transcribed conversation. However, even the clearest recordings have utterances that are hard to hear properly. From within a conversation, a participant would indicate to the speaker that they had not heard and ask for the misconveyed information, but obviously this is impossible when dealing with recorded speech. So the data is almost certain to be incomplete. Even going back to the speaker himself and asking him what he said will not guarantee filling in all the gaps in a transcription; in the case when the flow of speech is interfered with in some way, the speaker in question may be unable to recall what he said. Such is the transient nature of speech.

There is a wide range of fine- to coarse-grained options for the representation of different aspects of speech in transcription. One can choose to add detail down to the phonetic level in order to represent as closely as possible the exact manner in which something was said, including indicators of prosodic features; or one can abstract away from the nitty-gritty to the extent of producing a cleaned up, edited version of the speech (without the interruptions, mistakes, false starts, etc. that typically characterise unscripted, spontaneous conversation). However, the more you add to the transcription and the richer the information transcribed, the

less easy it is to read and the more time consuming it is to process, both manually and automatically. It is generally true that, what one gains in accuracy of representation one loses in readability, ease of processing and space. One would ideally like to be able to retain as much information about how an utterance is said as possible in a transcription, as one never knows whether features of language that appear insignificant or irrelevant, might not turn out to be more important than we suspected.

For my own research, I shall be looking at such a coarse-grained phenomenon, that I can afford to use a transcription abstracted at quite a high level from the actual sounds made. I retain interruptions, repetitions, and some contracted forms of words (such as 'gonna') as they are uttered; I also include fillers ('um' and 'er') and acknowledgements ('mm').

So, to recap, recording and collecting one's own data for research is difficult because:

(1) Transcription is extremely time-consuming (a ten-minute conversation may take more than a day to transcribe).

(2) Even the clearest recordings have utterances that are hard to hear properly, and are therefore uninterpretable when transcribed.

(3) There are issues of privacy that are at odds with recording natural, spontaneous conversation (which arguably can only be done when the recording is carried out covertly).

(4) There must not be more than three participants, otherwise there is likely to be more than one conversation that takes place at the same time (unless the conversational situation has a high degree of structure, such as in a meeting for example).

(5) Knowing the context of the conversation matters when one is interested in studying the underlying intentions of the participants and the structure of the dialogue as a whole. Without knowing the context, you risk not understanding the conversation.

(6) The type of setting of the conversation also matters. There are all sorts of interaction types, depending upon a large number of different factors: gender, age, authority, race, social class, etc. I am proposing to study the most generic type of conversation, in order to try to abstract away from some of these variables and uncover a generalised structure of interaction. Early research discussions about where I might obtain data included the possibility of using television broadcasts. Unfortunately, much of what happens on television is either scripted, or pre-planned, or so rigidly structured as to be of little use. In the case of interviews for instance, not only is control of the conversation firmly with the interviewer, but much of what is said has been rehearsed beforehand. Because of the expectations of an interview scenario, question-answer sequences of acts will dominate the conversation. So, studying this genre of 'conversation' would produce unbalanced results.

Aside from the difficulties of representing orthographically the spoken conversations themselves, I will now take a look at why I felt that the speech act annotation schemes that have already been developed by others are insufficient in some respect or other, and how this led to the evolution of the model I present here. (To some extent I shall be covering similar ground to that which I have already covered in Section 6.2.5, where I made a critical appraisal of the application of some of these schemes.)

7.3 Speech/Dialogue Act Annotation Schemes

There are an increasing number of competing speech act labelling and classification schemes in the computational linguistic community. This fact reflects the enormous interest that has been generated in recent years for analysing language at a functional level, with the idea of aiding in the design of human-computer dialogue systems. Often in the literature relating to dialogue management these are called *dialogue* acts rather than *speech* acts because the medium of interaction can be either written or spoken.

Most of the dialogue act schemes that are developed today focus on specific domain and task-oriented interactions; subsequently they have a tendency to be rather shallow and cross-grained (a mixture of too fine and too coarse). These schemes have been designed to be narrow for the most part to reduce the number of possible annotations, and to increase the rate of analysis for the NLP applications for which they have been developed. However, because there is a specific application in mind from the design stage, in all but a very few schemes, generality of use is lost. An overall comparison of features of a variety of different dialogue act annotation schemes is shown in Table 7.4. This is adapted from the MATE coding schemes, deliverable D1.1, Klein et al. (1998).

There have been a variety of attempts to come up with a definitive set of dialogue act labels. The main problems are that each different scheme was influenced by the theory behind it, and also by its application and use (i.e. what the developers wanted to use each scheme for). The results seem to be that each scheme differs from the rest just enough so that a direct mapping from one scheme to another is impossible. A comparison of the different dialogue act schemes shows the expected similarities between those that are used for equivalent or comparable domains; surprisingly, sometimes schemes that have totally different functional backgrounds show a considerable overlap. There is no way to tell however, whether this is because the design of one scheme has had an influence on the design of another. Trying to map all the representations onto each other, or amalgamate the various acts left unaccounted for in some schemes, into one overarching, generalised scheme causes problems at a theoretical level, because phenomena that should be considered at different levels of abstraction are conflated.

Schemes		**ALPARON**	**CHAT**	**CHIBA**	**COCONUT**	**CONDON & CECH**	**C-STAR**	**DAMSL**	**FLAMMIA**
Coding Manual		✓	✓	✓	✓	✓	✓	✓	✓
Annotators	*Number*	3	Very large	10	2	5	5	4	7
	Expertise	Expert	Expert	Expert	Expert	Fairly expert	Expert	Expert	Trained
Information about Annotated Dialogues	*Dialogues*	500	(160MB)	22	16	88	230	18	25
	Languages	Dutch	Many	Japanese	English	English	English, Italian, Japanese, Korean.	English	English
	Participants	2	2	2	2	2	2	2	2
	Task Driven	✓	×	✓	✓	✓	✓	×	✓
	Application Orientation	✓	×	×	✓	✓	✓	×	✓
	Domain Restriction	Directory Enquiries	×	Directions/ Appointments/ Travel	Furnishing Rooms Interactively	Transport	Travel	×	Directory Enquiries
	Activity Type	Information Extraction	Chatting	Cooperative Negotiation/ Problem Solving	Cooperative Negotiation	Cooperative Negotiation	Cooperative Negotiation	Several	Information Extraction
	Human/Machine Participation	Machine Mediated	Non-Machine Mediated	Non-Machine Mediated (?)	Machine Mediated (Computer)	Non- & Machine Mediated	Human-Human	Human-Human	Machine Mediated
Evaluation		77% Agreement	×	$0.57 < \alpha^{11} < 0.68$	✓	91% Agreement	×	$\kappa = 0.56$	$\kappa = 0.6+$
Mark-up Language		Own	Own	SGML-like	Nb	Nb	✓	DAMSL	✓
Annotation Tools		OVR coder	✓	Modified DAT	Nb	Nb	×	DAT	✓

(Continued overleaf…)

[11] The α (alpha) value (Krippendorff 1980) is calculated as: $\alpha = 1 - (D_O / D_E)$, where D_O describes the observed disagreements and D_E describes the expected disagreements.

Schemes		**JANUS**	**LINLIN**	**MAPTASK**	**NAKATANI**	**SLSA**	**SWBD-DAMSL**	**TRAUM**	**VERBMOBIL**
Coding Manual		✓	✓	✓	✓	✓	✓	✓	✓
Annotators	*Number*	4	4	4	6	7	9	3	3
	Expertise	Expert	Expert	Expert	Naive	Expert	Expert	Expert	Naive
Information about Annotated Dialogues	*Dialogues*	Many	140	128	72	100	1155	36	1172
	Languages	English	Swedish	English	English	Swedish	English	English	English, German, Japanese
	Participants	2	2	2	1	2 (?)	2	2	2
	Task Driven?	✓	✓	✓	✓	✓	×	×	✓
	Application Orientation	✓	✓	✓	✓	✓	×	✓	✓
	Domain Restriction	Business Appointments	Travel/ Transport	Directions	Instructions	Courtroom Interactions	×	×	Appointments
	Activity Type	Cooperative Negotiation	Information Extraction	Problem Solving	Teaching/ Instruction	Several	Several	Cooperative Negotiation	Cooperative Negotiation
	Human/Machine Participation	Human-Human	Non-Simulated	Non-Machine Mediated	Non-Machine Mediated	Non-Machine Mediated	Machine Mediated (Telephone)	Non-Machine Mediated	Non-Machine Mediated
Evaluation		89% Agreement	97% Agreement	$\kappa = 0.83$	×	✓ (Not published)	$0.8 < \kappa < 0.84$	✓ (Not published)	$\kappa = 0.84$
Mark-up Language		Own	Nb	Own SGML based	Nb	Own	Variant of DAMSL	Nb	Own
Annotation Tools		×	Nb	Own	Nb	TRACTOR	×	Nb	AnnoTag

Table 7.4 A comparison of different dialogue act annotation schemes.

Out of the many current dialogue act schemes, I will now look in detail at the two that, in my opinion, show the extremes of what is available. The characteristics of the VERBMOBIL and (SWITCHBOARD) SWBD-DAMSL codes plainly show why developing an appropriate scheme is problematic, and also highlight the various criticisms I want to level at previous work.

I have already made a critique of the application of these two schemes to practical problems (in Section 6.2.5). I now briefly consider why the schemes themselves fail, at a more fundamental level, to provide adequate coverage of speech act theory.

7.3.1 The VERBMOBIL Scheme

The European project VERBMOBIL was set up with the aim of developing machine translation in a fairly restricted domain – originally that of arranging business meetings, although in the second release the domain was widened to arranging travel plans also. English and German corpora were collected, annotated and software trained to recognise the dialogue acts being performed, as I have discussed earlier.

The scheme works well for the limited purpose for which it was created, but is too narrow-scoped to be easily expanded to cope with any other type of transactional dialogue, let alone with general conversation.

A more serious criticism perhaps (although also linked to the observation of specificity as mentioned above), is that the designers of the VERBMOBIL dialogue act scheme appear to have conflated the semantic and pragmatic levels of interpretation by including the content of an utterance in the dialogue act label itself, such as REQUEST_CLARIFY. Although this act is subsumed by an upper level dialogue act REQUEST, for reasons of shortcutting the identification of following acts, researchers have added the content type to the label itself. From the point of view of obtaining immediate results this is difficult to fault, as there is little doubt that the content does to some extent influence the identification of the speech act being performed.

However, it is unclear to me that you can admit such merging of levels from a theoretical standpoint. The fact that the request is for a *speech* action, should not distinguish it from a request for any other kind of action, as there are an endless list of such possible labels: REQUEST_INFORM, REQUEST_RENEW, REQUEST_LEAVE, etc. The scheme is obviously not generic enough because the ratified speech acts would proliferate infinitely. Thus the expressivity of the scheme is called into question, as there are possibly as many of these specific dialogue acts as there are utterances!

The problem of the endless propagation of speech acts is clearly highlighted in the work of Reithinger and Klesen (1997), in their statistical model of dialogue act recognition for the specific task of arranging business appointments. The dialogue act decoding program can

recognise the following as valid acts: SUGGEST_SUPPORT_DATE, FEEDBACK_RESERVATION, ACCEPT_LOCATION, etc.

7.3.2 The (SWITCHBOARD) SWBD-DAMSL Scheme

At first I had intended to amalgamate the dialogue acts in the SWBD-DAMSL scheme with the acts I identified in the model developed during earlier work (Schiffrin 1995). I thought that this would be appropriate considering that it is the only scheme available that has been applied specifically to general conversation. However, it soon became obvious that there were a number of problems with simply merging the two schemes. These problems stem mainly from a number of criticisms that can be levelled at the SWBD-DAMSL annotation scheme itself:

(1) *Speaker/Hearer distinction*: There is a problem with annotating utterances that are produced with the intention of being identified as one kind of act, but being interpreted by the hearer as another, against conversational expectation. For example, in the case of an or-question that is either interrupted, or unfinished for some other reason, and which is taken by the hearer/addressee as a YES-NO-QUESTION. E.g.:

A: Did you bring him to a doggy obedience school or –
B: No –

This would not be a problem in my representation, as the conversational models of each participant would reflect these differences in intention and interpretation.

(2) *Only two participant dialogue*: This scheme is constrained to a maximum of two participants, probably due to its development on telephone conversation. If one of the speakers begins to talk to a person other than the person at the other end of the line, then this utterance is labelled 3RD-PARTY-TALK at the communicative status level. I wonder whether the scheme is actually usable on more than two participants or not. It could be that external talk was excluded simply on the grounds of being unable to capture the other person's responses (which would thereby introduce an inconsistency at the probabilistic level of analysis for the purposes of *n*-gram and hidden Markov modelling). This is unclear from the description presented by Jurafsky et al. (1997a).

(3) *Conflation of different theoretical levels of mark-up*: This is a twofold criticism. The first is a problem with the way labels can proliferate for each utterance. There are five dimensions accounted for within the scheme: *communicative-status*, *information-level*, *forward-communicative-function*, *backwards-communicative-function*, and *other*. Each of the dimensions can 'bid' for the utterance, which in fact can be multiply tagged. There are a dizzying number of possible combinations. Because of the complexity of the original

SWBD-DAMSL scheme, it is difficult to see how it can ever be possible to use it in order to annotate dialogue reliably. The flattened SWBD-DAMSL scheme suffers from a similar problem, despite efforts at simplification. It is impossible to tell the precise definition of a speech/dialogue act at all according to this scheme, as so many different theoretical levels are merged and blurred together (for example, the conversational move level with the dialogue act level).

The second criticism is the way even within a dimension different linguistic features are confused. Dialogue acts were first developed as purely pragmatic markers – the SWBD-DAMSL tags include syntactic information as well. As an example, questions that are made declaratively are explicitly marked as a separate act, yet these still function as *questions* at the pragmatic level.

7.3.3 Theoretical Distinctions

Both the VERBMOBIL and SWBD-DAMSL schemes fall under the same criticism that they have no firm theoretical foundations for their choice of dialogue acts. It is impossible to tell whether there are any acts that are not covered by their scheme, because we do not know the criteria for the inclusion of the dialogue acts defined. We have no means of knowing whether the dialogue acts are comprehensive.

From the study of the various schemes available, and especially from the consideration of the last discussed, SWBD-DAMSL, a number of features that are relevant for the application of a dialogue act scheme to spoken, general conversation begin to stand out. These can be listed as follows:

(1) *Generic vs. specific?*: Has the scheme been designed for use in a specific situation, with a specific application in mind?

(2) *Open vs. domain restricted?*: Has the scheme been restricted to a certain domain of application (i.e. a closed knowledge base of facts and goals)?

(3) *Non-goal-driven (general) vs. task-oriented conversation*[12]*?*: Is the scheme used in transactional dialogue only?

(4) *Observed vs. theoretically structured annotation scheme?*: Have the dialogue act label names been chosen for theoretically structured reasons, or have they been allocated to repeated, observed phenomena in data?

---

[12] As I have noted before, arguably all types of dialogue are goal-driven to some extent.

(5) *Speaker/hearer distinction?*: Is there a distinction made between the speaker's production and the hearer's interpretation of the dialogue act?

(6) *Acoustically informed annotation process?*: Has the annotation process involved expert annotators with access to the original audio recordings?

(7) *Is any number of speakers covered by the annotation?*: Does the scheme accommodate and account for any number of speakers in a conversation?

(8) *Deals with long distance dependencies and discontinuities in the data?*: Does the dialogue act scheme deal with non-sequential dependencies, and discontinuous dialogue act analyses?

(9) *Distinction between different levels of abstraction?*: Are the different levels of abstraction of utterances distinguished? By 'levels of abstraction' I mean whether the act being performed is in fact not a pragmatically interpreted dialogue act at all, but an act (often conventionally) performed for social or turn-management purposes. These are often treated as the same type of entity in annotation schemes, but I shall argue that they are not. I shall be discussing the problem of distinguishing between different levels of abstraction in Chapter 9.

(10) *Semantic/pragmatic distinction?*: Is the content of the utterance excluded from the dialogue act type? For example, 'location' in VERBMOBIL's ACCEPT _LOCATION should be part of the semantic content of the utterance, whereas ACCEPT is the generic pragmatic speech function.

(11) *Syntactic/pragmatic distinction?*: Is the syntactic form of the utterance separate from the speech act type? For example, SWBD-DAMSL's DECLARATIVE YES-NO-QUESTION tag has the syntactic information about the form of the utterance, which performs the generic pragmatic speech function YES-NO-QUESTION, coded within the tag.

If we compare (in Table 7.5) two of the most widely used schemes available with each other, we can see how they match up to this list of criteria:

	VERBMOBIL	SWBD-DAMSL
(1) Specific/Generic?	Specific	Generic
(2) Open/Domain Restricted?	Domain	Open
(3) Task-Driven/General?	Task	General
(4) Observed/Structured?	Structured	Observed
(5) Speaker-Hearer Distinction?	×	×
(6) Acoustically Informed?	✓	×
(7) > 2 Speakers?	×	×
(8) Discontinuous Data Dependencies?	×	×
(9) Levels of Abstraction?	×	✓/×
(10) Semantic/Pragmatic Distinction?	×	×
(11) Syntactic/Pragmatic Distinction?	✓	✓/×

Table 7.5 Comparing the theoretical distinctions of VERBMOBIL and SWBD-DAMSL dialogue act annotation schemes.

The theoretical distinctions that have been brought out here are, in my opinion, fundamental for the development of a dialogue (or speech) act annotation scheme that is comprehensive and fully grounded in pragmatic theory. This is not to say that syntactic and semantic features have no part to play in the identification of speech acts in discourse; however, I do think that, rather than being subsumed by the overall act, syntactic and semantic influences on utterance understanding are separate qualities that contribute to the underlying meaning. This is what I hope to show in the remaining chapters of this dissertation.

In this chapter, I have explained why it is of interest to study conversational speech, described the drawbacks of using available spoken corpora, the pitfalls of data collection, and the deficiencies of current dialogue act annotation schemes. In the next chapter, I will start to suggest a theoretically rigorous method for capturing the functionality of dialogue (or speech) acts.

Chapter 8

Language Use and Context

> 'Come, we shall have some fun now!' thought Alice. 'I'm glad they've begun asking riddles. – I believe I can guess that,' she added aloud.
>
> 'Do you mean that you think you can find out the answer to it?' said the March Hare.
>
> 'Exactly so,' said Alice.
>
> 'Then you should say what you mean,' the March Hare went on.
>
> 'I do,' Alice hastily replied; 'at least – at least I mean what I say – that's the same thing, you know.'
>
> 'Not the same thing a bit!' said the Hatter. 'You might just as well say that "I see what I eat" is the same thing as "I eat what I see"!'
>
> (*Alice in Wonderland*, Lewis Caroll)

One of the principal premises of this dissertation is that language can be analysed as a series of actions: that language is used to perform activities by means of speech acts. In Chapter 6, I looked at two related, but different, approaches to the computational recognition of speech acts in discourse: one based on the identification of a speaker's intentions in producing an utterance (an inferential paradigm), and the other based on the hearer's conversational expectations (a structural paradigm). In the remaining chapters of this dissertation, I shall argue that speech acts (at the level of abstraction that I shall define them) primarily indicate, by their very nature, the *structure* of communication; I maintain that communicative inferences are only derived secondarily. The model that I have developed therefore mainly takes a structural approach to speech act recognition, and provides a framework for an inferential model of conversational meaning. I justify this theoretical stance for the following two reasons:

(1) Whilst neither one of these methodologies is sufficient in itself, the structural approach provides the background from which a speaker's intentions can be understood. In other words, it is the foundation upon which to build an interpretation of an utterance. The point is, without a solid foundation, the constructed interpretation of an utterance within a conversation will not be sturdy or reliable.

(2) It is the structural approach that lends itself best to computational modelling, especially when dealing with generic models of interactional behaviour. This is because under this paradigm, the analysis of an utterance is less reliant on background knowledge and logical inferences, and more so on the pattern of acts to which the current utterance belongs. It must be stressed again though that this approach is incomplete on its own, and is therefore not standalone.

The theory of language understanding that I present here has much in common with conversational game theory (see Section 6.2.2) and context change theory (see Sections 6.2.3 and 9.1.1). Although the model I expound is related to other structural analyses, it is also significantly different. For one, an account is given of observed phenomena in spoken, general conversation (especially where the distinction between speaker and hearer status is concerned). For another, a deeper analysis of conversational structure than has hitherto been explored by other researchers is proposed, which allows for the complex inter-speech-act dependencies that are to be found in naturally occurring conversation.

Apart from speaker-oriented and hearer-oriented models of communication (which both essentially view the speaker's role as active and the hearer's role as passive), Clark (1996) suggests a third view of language use. In common with other researchers (particularly proponents of conversational game theory, such as Hulstijn 2000, and those modelling communication for computational agent collaboration, such as Wooldridge and Jennings 1994a and 1994b, Grosz and Kraus 1999), Clark views language as essentially a form of **joint activity**. Conversation, he argues, which is performed between two or more people, has many of the features found in other human joint activities[1]. The view of language as a joint activity implies that the hearer's role is just as much active as that of the speaker. In Section 8.2, I shall explore some of the general properties of activities that are taken on by two or more people, and compare them with the properties that are displayed in conversation. This will then lead to the introduction of the model I wish to propose in this dissertation. Firstly though, we shall take a look at why it is incorrect to assume that, to understand the function of an utterance in conversation, one need only take the speaker's viewpoint into account.

8.1 Speaker Meaning and Hearer Understanding

As we have already discussed, although traditional approaches to speech act theory do take some account of the hearer's role in understanding (e.g. Leech 1983), the importance of the hearer's uptake of the speech act being performed is not always fully explored.

---

[1] Although I have adopted and adapted some of the ideas presented in this chapter from Clark (1996), who outlines an excellent theory of language use as joint activity, the main points that I wish to draw out here are not part of Clark's thesis. Indeed, in Section 8.1, I will dismiss Clark's rather strong version of joint action in favour of a solipsistic approach.

When one looks at real conversation, it becomes clear that in order to perform a speech act successfully, at least one accompanying speech act (a response of some kind) is generally required from the hearer. Austin noticed this feature of speech act performance; he has the following points to note when he gives the example of a case when he has promised to return some money to a friend by saying 'I'll pay you back tomorrow':

> It is obviously necessary that to have promised I must normally (A) have been *heard* by someone, perhaps the promisee; (B) have been understood by him as promising. If one or another of these conditions isn't satisfied, doubts arise as to whether I have really promised, and it might be held that my act was only attempted or was void. (Austin 1962: 22)

Austin also observes that:

> One of the things that cause particular difficulty is the question whether when two parties are involved 'consensus ad idem' is necessary. Is it essential for me to secure correct understanding as well as everything else? (ibid.: 36)

I shall return to this last question about whether what Austin calls 'consensus ad idem' is necessary in Section 8.1.3 when we look at an example conversation; I shall argue that exact understanding is not strictly de rigueur. It is clear that some form of uptake, or at least some sign that the intended act has been recognised, is needed for its successful performance. This is often accomplished in spoken discourse by means of feedback and paraphrase, or more commonly, as I will argue, by the use of an appropriate speech act in response.

If we look at Austin's interpretation of the different levels of acts that go on in an utterance such as 'Please sit down', we get the following list of acts:

Phonetic act	I am producing the noises that constitute 'Please sit down'.
Phatic act	I am uttering the words *please*, *sit*, and *down*.
Rhetic act	I am using the words *please*, *sit*, and *down* with a certain sense and reference.
Locutionary act	I am saying to you 'Please sit down'.
Illocutionary act	I am asking you to sit down.
Perlocutionary act	I am trying to get you to sit down[2].

---

[2] As an aside, I have a criticism of Austin's list of acts performed in speech. The definition of the perlocutionary act seems wrong somehow – surely if I actually get you to sit down by producing this utterance, then that is the perlocutionary act? The definition given here is the recognition of the underlying intention that I might have if uttering this sentence (and therefore of great interest to me, because it is this that tells the hearer what to do with the content of the utterance), but it is not a perlocutionary *act*. This phrase may well be, however, the intended perlocutionary *effect*.

Note that all of the acts in this list are speaker-oriented and show no consideration of the role of the hearer within a conversation. Clark (1996) claims that the definition of acts given by Austin is not representative of what actually occurs in communication. In communication, speakers *execute* a behaviour for their addressees, who then *attend* to that behaviour in turn; this therefore constitutes a joint action, with both parties participating.

8.1.1 Speech Action Ladder

Clark rejects the hierarchy of acts suggested by Austin in favour of what he calls a **speech action ladder**, which reformulates Austin's set of acts, taking the hearer's role into consideration (see Table 8.1, taken from Clark 1996:153).

Level	Speaker's view	Hearer's view
4	Proposal	Consideration
3	Signalling (meaning)	Recognition (understanding)
2	Presentation	Identification
1	Execution	Attention

Table 8.1 Language joint action ladder

So, if we look at Austin's example of the utterance 'Please sit down' again, but from the perspective of an action ladder, we get the following kind of analysis:

Level 1: The speaker executes the sounds 'Please sit down'. The hearer attends to the speaker's execution. (= Phonetic and Phatic)

Level 2: The speaker presents the words 'Please sit down'. The hearer identifies the speaker's presentation. (= Rhetic and Locutionary)

Level 3: The speaker signals the request that the hearer sit down. The hearer recognises the speaker's signal. (= Illocutionary)

Level 4: The speaker proposes that the hearer sit down. The hearer considers the speaker's proposal. (= Perlocutionary)

All the actions on the ladder take place simultaneously during an utterance so that one might be tempted to say that one cannot really distinguish between them. But we can tell that these are in fact not one act by the dependency that the successful completion of one level has on the successful completion of another (Clark 1996). So, the speaker executes an articulation of sounds *in order to* present a message, which he does *in order to* signal that he is performing some speech act, which he does *in order to* propose some action. The chain of causality from the bottom level up to the top leads to the idea of *upward completion*. This means that it is only

possible to complete the overall action when each sub-level is fulfilled satisfactorily. To correctly perform the top-level action, no stage can be missed out.

We can reverse the order of the chain, but the dependency is still uni-directional, through Levels 1 to 4. The speaker can be said to be proposing some action *by means of* signalling that he is performing some speech act, which he is doing *by means of* presenting a message, which he is doing *by means of* executing an articulation of sounds. This gives a chain of *downward evidence* through the levels, so that when we know that one level is complete, we have the necessary evidence that all levels below it are complete also.

Knowledge of the chain of upward completion gives the hearer a means of reconstructing the speaker's motives and goals, while downward evidence is used by the speaker to span from a top level goal to the production of the sounds that should be used to attempt to achieve it. In essence, both comprise the same set of relations, but represented from different points of view.

8.1.2 Level Failures

These two characteristics of upward completion and downward evidence are important not only in determining what an utterance means, but also when it comes to identifying where a conversation has gone wrong. It is when there is a block to the completion of one of the action levels that communication problems occur, whether they are actually perceived by the speaker or hearer, or not. Communication can fail at any point on the speech action ladder, which is why a speaker in a conversation requires feedback from his audience – to reassure himself that the top-level action has been interpreted successfully. Here are some of the consequences of failure at the different levels:

(1) *Attention*: If I enter the house and shout, 'Hello, is anyone home?', and there is no reply, even though my brother, Bob, is upstairs playing the piano with his earphones on, then my attempt to get someone to identify my shout as a question that expects a response at Level 4 of interpretation, has failed. I cannot infer that there is no-one in the house simply because there is no reply; or if I do, it is a mistaken inference, because I have assumed that my utterance can be heard. In this case I have not even succeeded in getting on the hearer's scale at Level 1.

(2) *Identification*: Note that in the example above, if Bob had shouted down, 'What?', my question about whether there was anyone in the house would have been answered, but not intentionally. The communication would not have got past Level 1 for the hearer – which is the recognition that there was someone in the house trying to say something. They have heard that there is a message, but not what that message is. This is a failure at Level 2 of the action ladder. It is analogous to if we were to be lost in a foreign city, and were to ask a

passer-by, 'Do you speak English?'. If the passer-by looks at us blankly, or says, 'No le entiendo' ('I do not understand you' – we are, it seems, lost in Buenos Aires), then we have failed to communicate, not because our hearer has not heard the message, but because he is unable to decode it.

(3) *Recognition*: Communication can fail at Level 3 when the signal is heard, and the words decoded, but the wrong inference is drawn about the meaning of the utterance presented. So for example, supposing I ask a friend, 'Have you got the time?', and he answers, 'Yes, I'm not that busy at the moment – what can I do for you?', but it turns out that in actual fact I was requesting to know the time, and an appropriate answer would have been, 'Yes, it's half past four'. I can tell that there is a problem, because the question I have asked legitimises an answer from the set of possible hours in the day. In this case my friend has been unable to recognise correctly my intention in producing the utterance, and his understanding only got as far as Level 2 (although interestingly enough, he thinks he has reached Level 4). Sometimes questions are wilfully misunderstood, and this becomes a source of humour (e.g., if my friend had answered my question, 'Have you got the time?' with the phrase, 'If you've got the inclination, baby!', I would still have failed to get my friend to respond as I intend at a level past Level 2, but would interpret this as mere sauciness on his side and expect a proper response to ensue).

(4) *Consideration*: If we take the same question again, 'Have you got the time?', and my friend were to answer, 'I'm sorry, I don't have a watch', this demonstrates a failure at Level 4. Although my friend has successfully understood that I'm asking him to inform me of the time, he is unable to take up my (implicitly) proposed action because one of the prerequisites for its performance is missing (i.e. that he has a means of telling the time at his disposal).

Of course these action failures are often not so simple, because it only takes one word to be misinterpreted, or to be uninterpretable for some reason, by the hearer for the entire action ladder to fail (because the completion of every level is crucial).

Notice also that these communication failures are all presented from the hearer's perspective. It is often difficult to ascribe 'fault' when it comes to misunderstanding; in the example given in (2) above, who is to blame for the failure to communicate – the speaker for speaking the wrong language, or the hearer for not understanding the correct language required to interpret the utterance? In some cases it is clear that the cause of miscommunication lies with the speaker, because he has made a mistake, e.g. by using the wrong word, or by not making the function of his utterance clear enough within the current context. However, although mistakes can be made in production as well as interpretation, in practice it is unimportant from where the error stems. A perceived slip-up will produce the same effect (i.e. usually that of initiating some sort of

repair sequence) regardless of who is responsible; an error that remains unperceived by all parties, while having an effect on the different mental models of participants in the current conversation, will generally have no serious consequences on the conversation itself. Note this is not to say that an unperceived misunderstanding might not have unfortunate repercussions for the participants. If we have arranged to meet and have understood different times for the proposed meeting, then we will fail to meet. But if the problem is unrecognised at the time of making the arrangement, we will come away believing different things, while also believing that we have understood each other.

It is partly because it does not actually matter in fact how a communicative act fails that I am led to reject Clark's view of language as a joint activity, in favour of a fourth paradigm that, as far as I am aware, has hitherto remained curiously overlooked in modern theories of communication. Before outlining my own suggestion for dealing with the distinction between speaker meaning and hearer understanding, I shall clarify the problem by looking, in Section 8.1.3, at a step-by-step example of conversational failure and its consequences, which is taken from real life.

8.1.3 Reference Failure and Model Failure

Although we have determined in Section 8.1.2 (at least) four ways that communication can fail, in fact generally these boil down to two types of failure in the end. Level 1 and 2 failures equate to what Grau et al. (1994) term input failure, which are those errors that are brought about either because there is a lack of hearing or understanding. I prefer to call such communication problems **reference failure**, because it may not just be an input failure on the side of the hearer, but an output error on the side of the speaker. As I have argued previously, the resulting lack of understanding is the same no matter where the original error lies. Level 3 and 4 failures[3] are encompassed by the term **model failure**, because they are recognised by a failure to fit the utterance into the hearer's structural or informational model of the conversation. So, reference failure is characterised by the hearer being unable to build a coherent semantic representation of the utterance, while model failure occurs when the hearer understands, or believes he understands the message, but is unable to integrate it into a coherent pragmatic model of discourse.

---

[3] Strictly speaking, a Level 4 action failure is not a communication failure at all, but a failure to achieve the goal of the communication. In fact, my research is not really concerned with the achievement of Level 4 actions; this is, in my opinion, more properly a point of interest for those working on inferential models. It is the identification of Level 3 actions, which equate to illocutionary acts, that is of most importance here.

It is often the case that reference failure will lead to model failure, especially if a participant is unaware that the reference failure has occurred in the first place (this is what happens in the example shown below in Conversation 8.1), and this is why the expectations we have about the kind of information that should follow each utterance is of paramount importance in successful communication.

Once a participant identifies a level failure, it is his responsibility to repair the communication if possible. Most often it will be the initiating participant who tries to reformulate his original utterance (there is in general a preference for self-repair of communication among human beings, for reasons of 'face' and power – Schegloff et al. 1977); however, sometimes the hearer will try to prompt this repair by saying something like, 'Did you mean *X*?'. Conversation 8.1 (which took place in 1997) shows a snippet of a conversation between the author and a male friend, Steve (this is not his real name). The conversation exemplifies the impact of reference failure and model failure.

	(1) **Steve**:	What are you doing on Friday night?	
①	(2) **Mandy**:	I'm going out with Claire and her sister.	
	(3) **Steve**:	Sarah…	②
	(4) **Mandy**:	No, Vicky.	
	(5) **Steve**:	Sarah…?	
	(6) **Mandy**:	Sarah? No, *Claire* and her sister, Vicky.	③
	(7) **Steve**:	Oh Claire! You said Sarah.	
	(8) **Mandy**:	No I didn't, I said Claire.	
	(9) **Steve**:	Oh.	

Conversation 8.1 Snippet of conversation between Steve and Mandy[4].

This conversation can be analysed in the following way:

[4] Although this example conversation was never actually recorded, I wrote it down immediately after it had taken place, precisely because it is such a good example of the kind of phenomena in which I am interested. Therefore, while it may not be word-for-word exactly as it was spoken at the time, it is as nearly correct as memory can be relied on after a lapse of a few minutes.

① **REFERENCE FAILURE**: After utterance (2), Steve has misheard 'Claire' as 'Sarah'[5]. At this stage he is unaware that the conversation has gone wrong, so asks for more information about Sarah in utterance (3).

② **MODEL FAILURE**: This in turn is misinterpreted by Mandy who thinks this is a request for clarification that the sister's name is Sarah, so denies this and informs that the sister's name is Vicky in utterance (4).

① This does not help Steve with his request for more information about Sarah, so he repeats his question in utterance (5), now aware that something is wrong.

③ **RECOVERY**: In utterance (6), Mandy at first does not understand the question (she has not mentioned Sarah, so this seems to break Gricean rules of co-operative relevance in conversation). She realises that there has been a reference failure (Steve has misheard something), corrects the original mistake then explicitly names the sister. This could be for two reasons: firstly, to state the information explicitly, and secondly, to explain implicitly her seemingly unco-operative behaviour in (4) when she fails to answer Steve's question (because she wrongly interpreted utterance (3) as a statement). In utterance (7), Steve also recognises the mistake, then claims the mistake was one of output (it was Mandy's fault). In utterance (8), Mandy denies output failure and claims input failure (it was Steve's fault). It is unclear whether there is any agreement as to 'fault', but this is irrelevant as the error is corrected nonetheless!

In fact, one suspects that the participants stubbornly retain their individual diagnoses of how the conversation went wrong, and therefore continue to believe different things about what was actually said. So, while both agree that Mandy intended to say 'Claire' in utterance (2), Steve still believes that Mandy actually said 'Sarah' and Mandy still believes that she actually said 'Claire'. This leads to contradictory beliefs in the participants' mental models, but this is not a problem in real life. However, for most computational models of belief, this would cause severe difficulties, because there is an assumption in modern approaches that the purpose of conversation is to unify the participants' mental models and to end up with what are commonly termed 'mutual beliefs' and 'common goals'.

I have shown the complete analysis of Conversation 8.1, given from both the speaker and the hearer's points of view in Table 8.2. There are three observations that I would like to make

---

[5] For the sake of fairness, I must admit that, for reasons that are too complicated to go into here, I was often confusing the name 'Claire' with the name 'Sarah' at the time that this conversation took place. I still maintain however, that I said 'Claire' and not 'Sarah'!

about this conversation (as it is analysed in Table 8.2) before suggesting a different way of modelling speaker meaning and hearer understanding.

The first point to note is the denseness of communication. The conversation itself took (roughly 10-20) seconds to complete, but the information conveyed and the inferences made by the participants are fairly complex (see Table 8.2). If all the information that is tacitly worked out by the participants had to be explicitly stated, communication would take orders of magnitude longer to perform. This observation comes by way of admiration for the simplicity and efficiency of the mechanisms we employ for communicating with each other.

The second point is that, although I have presented the analysis of the conversation from both my own and Steve's point of view, and as far as I am able to ascertain, having interviewed Steve concerning his interpretation of this conversation retrospectively, they are correct according to both participants, even so there is no guarantee that I have not added or missed out extra meaning than there actually was. This post-analysis of what went wrong and what we thought at the time of the conversation is still fairly unreliable in the sense that we will almost certainly never be able to capture our thoughts at source. So, while I believe that the interpretation I have presented is an accurate one, that is representative of the way we communicate with each other and the way we recover from errors, this itself is a highly subjective opinion.

The third point (which is related to the second) is that, we can clearly see from the example as it is decomposed in Table 8.2 how the speaker and hearer can in actual fact interpret different things from the same utterance. This is not an uncommon occurrence in conversation (see discussion in Chapter 2). Now we can understand why taking the approach of assigning a single meaning to every utterance performed by a speaker is not at all representative of what happens when people interact and therefore not sufficient when trying to model what happens in real conversation.

If one does not take both the speaker's and the hearer's perspective into consideration when designing a computational model of speech act interpretation, then one cannot build a robust spoken language system. The unreliability of the message, in terms of the errors in speaker output and/or in hearer input, means that complex, yet dependable methods and strategies of recovery must be in place for communication to take place smoothly. Not only that, but we must also be able to allow for people's models of the same conversation to 'come apart'; they are not identical, and, I believe, never can be.

	Speaker Means	Hearer Understands
(1)	**Steve**: I am asking Mandy what she is doing on Friday night.	**Mandy**: Steve is asking what I am doing on Friday night.
(2)	**Mandy**: I am telling Steve that I am going out with Claire and her sister ('on Friday night' is assumed because I am answering Steve's question).	**Steve**: Mandy is telling me that she is going out with Sarah and her sister ('on Friday night' is assumed because she is answering my question).
(3)	**Steve**: I am asking Mandy to tell me who Sarah is because I do not know which Sarah she means. Who is she?	**Mandy**: Steve is checking to see whether Claire's sister's name is Sarah. While I referred to Claire by name, I left Claire's sister unspecified; this was because, though Steve has met Claire, he has never met her sister.
(4)	**Mandy**: I am correcting Steve's mistaken assumption that Claire's sister's name is Sarah, and informing him that her name is actually Vicky.	**Steve**: Mandy is telling me that we are not talking about a girl called Sarah (and her sister), but about one called Vicky. This is confusing because I did not hear "Vicky", I heard "Sarah".
(5)	**Steve**: I am repeating my question as in (3) above, this time emphasising by my intonation that it is a question.	**Mandy**: Steve is asking me a question about Sarah, possibly asking me who she is from the tone of voice he is using. However, I have not been talking about a girl by the name of Sarah…
(6)	**Mandy**: I am repeating my act of informing in (2), this time clearly specifying the names of all the people to whom I am referring. I suspect that Steve has misheard what I said, and I include the full information, not only to be clear, but also to explain my inappropriate answer to Steve's question in (3).	**Steve**: Mandy is repeating the information she gave me in (2), only this time she is replacing the name Sarah with the name Claire. So either she said the wrong name, or I misheard what she said.
(7)	**Steve**: I am telling Mandy that I now understand that we have been talking at cross purposes and that we should have been talking about Claire; I am also telling her that she said the wrong name, Sarah, instead of Claire.	**Mandy**: Steve is telling me that he understands that we should have been talking about Claire, and he is asserting that I said Sarah by mistake.
(8)	**Mandy**: I am denying that I said the wrong name and asserting that I did indeed say the name Claire (by implication this also asserts that Steve has not heard my utterance correctly – after all, the names Claire and Sarah are fairly similar).	**Steve**: Mandy is denying that she said the wrong name and asserting that she did indeed say the name Claire (by implication this also asserts that I did not hear her utterance in (2) correctly).
(9)	**Steve**: I am acknowledging the error (in a non-committal way where the laying of blame is concerned).	**Mandy**: Steve is acknowledging the error (in a non-committal way where the laying of blame is concerned).

Table 8.2 Example of speaker meaning and hearer understanding coming apart.

8.1.4 Rejecting the Joint Activity Hypothesis

Although Clark (1996) argues cogently and convincingly for his theory of joint activity, he makes too strong an emphasis on the word 'joint'. We can clearly see that while we might aspire to a joint activity in communication, in reality we always maintain our own perspective of a conversation. We can highlight this distinction in approaches by looking at the case of someone speaking to himself. Clark says that what happens in this state of affairs is that we pretend to have someone's attention; but surely really we are taking on the roles of both speaker and hearer at once? Speaking to one's self can never go wrong in the same way as speaking to another person can (unless perhaps we are schizophrenic), because rather than having two models of the conversation on the go at the same time, there is just one, uncontroversial representation. In fact I believe that the same can be said for all so called joint activities, as Clark defines them, although most are less prone to ambiguous interpretation than speech is, because generally the context is purely a physical one. It is this observation that may explain why it is so singularly unsatisfying playing a game (say chess for example) on one's own. It is the surprise of the unexpected move that provides the amusement of the game. Consequently I believe nearly all language use is an elicitation of response from our fellow human beings.

Representing Clark's theory of joint activity computationally, would pose many problems. While there may very well be joint actions going on in real life, it would take a god's-eye view to be able to model this. In other words, we are not wholly aware of when we achieve 'jointness' and when we do not. We ourselves only ever get one point of view. We take on the role of speaker and listener in our turn, and can only judge the effect of our utterances on a hearer by the way we have learned to react ourselves to a perceived type of utterance. We cannot read each other's minds – we are one-channel beings and our mental representations of a conversation (in themselves fairly unreliable and subjective) can only be consulted by the same means as they are created. It is for this reason that I am uncomfortable with the word 'joint' or 'mutual' or 'common' when describing knowledge or actions represented by the speech of human beings; I think the most one can claim is that we perform co-ordinated actions.

It is for these reasons that I argue for a solipsistic[6] approach, to model the mental states of one person who constantly changes his role (and his perception model) from speaker to hearer. This, I contend, is a more psychologically valid viewpoint than assuming that we gradually acquire exactly the same information; after all, we have no access to another's internal mental states *except* by what he says (and does). I call this the **co-ordinated** (to replace the word 'joint') **activity** approach, because the word 'co-ordinated' reflects the fact that there are two

---

[6] Solipsism is essentially the philosophy that one can only ever know one's own conscious, subjective reality (and arguably even that knowledge is prone to error).

types of behaviours, or roles, in any activity that is taken on together by human beings, but that these are only ever analysed from a single, role-changing perspective.

The difference between the joint and co-ordinated activity approaches are demonstrated diagrammatically in Figures 8.1 and 8.2.

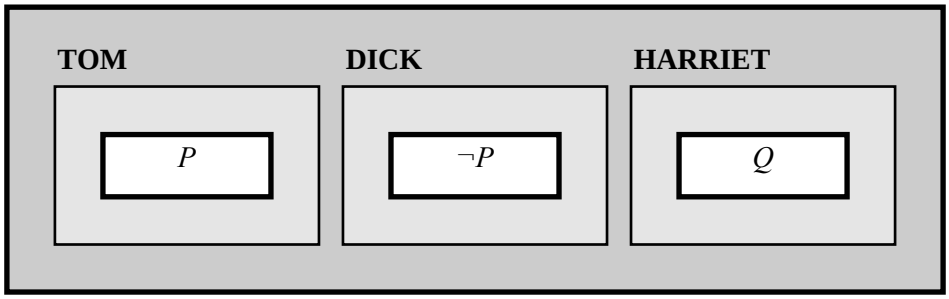


Figure 8.1 The joint activity approach.

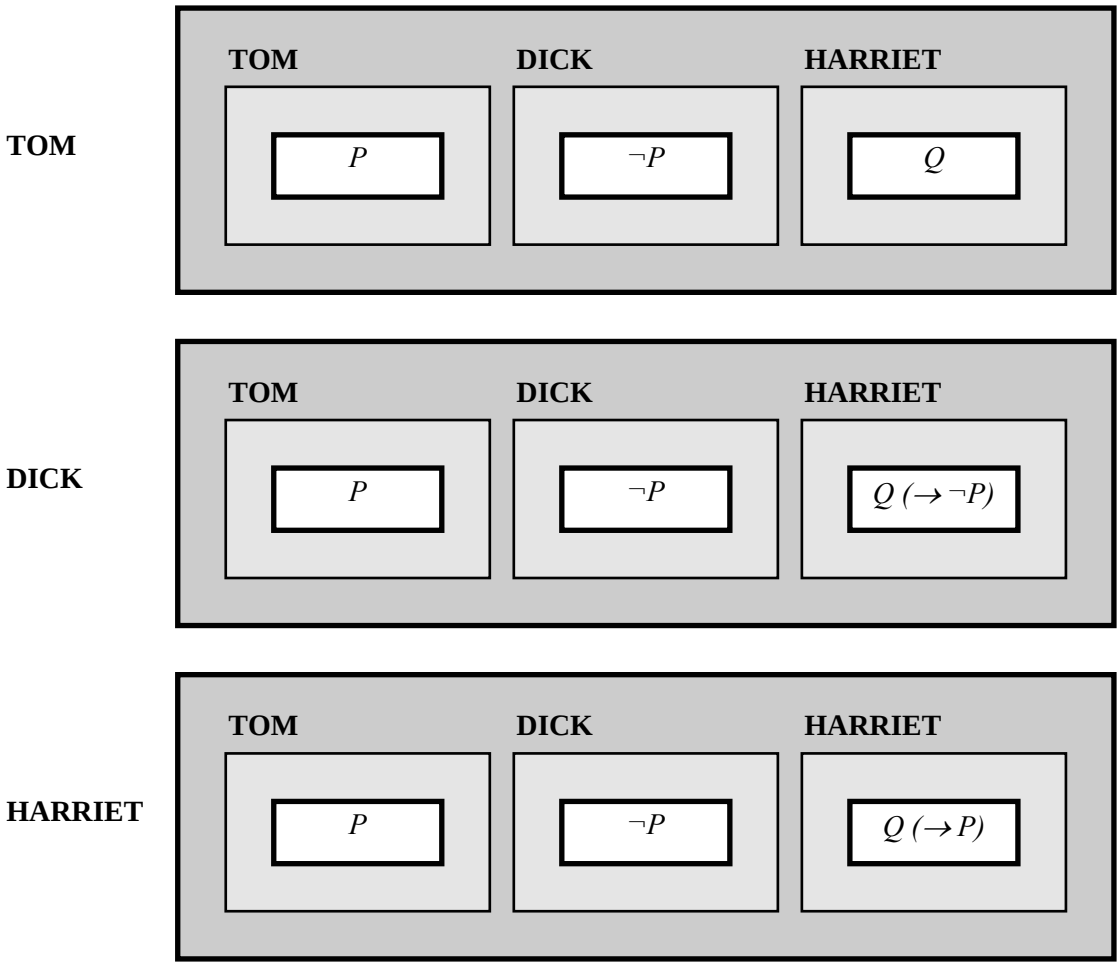


Figure 8.2 The co-ordinated activity approach.

Although the joint action perspective takes the role of the hearer into consideration, it is only a two-dimensional model that assumes that all utterances are public events that are uniquely interpretable by the participants. So, in the model shown in Figure 8.1, it is clear to all that TOM has said that *P*, DICK has said that ¬*P*, and HARRIET has said that *Q*.

In Figure 8.2 we can see that the co-ordinated action perspective offers more of a three-dimensional model of language. Of course this is a god's-eye view of the state of the conversation, but we can observe a number of interesting differences in each individual's model of the propositions that have been made in the conversation so far.

Let us look specifically at HARRIET's proposition, *Q*, in the diagram. Let us also assume that she was the last person to have spoken. Her utterance of *Q* was meant by her to indicate support of TOM by agreeing, by implication, with his proposition *P*. Unfortunately however, TOM does not know that $Q \rightarrow P$, and so is unable to interpret her intended meaning correctly. As for DICK, he believes that $Q \rightarrow \neg P$, so therefore also misinterprets HARRIET's proposition (as one of agreement with him and disagreement with TOM), but for different reasons to TOM. All three participants end up with different interpretations of the conversation (specifically concerning the function of HARRIET's utterance), but they will only find this out from the subsequent conversation, if at all.

Notice that, in Figures 8.1 and 8.2, I have omitted details about the order in which these propositions are carried out, and about the speech act interpretations made by the participants in order to achieve this state of affairs. In fact the diagrams are massively oversimplified; utterance interpretation in truth relies heavily on a four-dimensional model – the sequence and order of utterances play a crucial part in understanding the function of a particular utterance in the conversational context. I shall return to this point in Chapter 9; for the present I simply wish to illustrate why the view of language as a joint activity does not work.

Although we only have immediate access to one current model of the conversational state (having only one of these points of view ourselves), we must also be able to reason about how an utterance could be interpreted differently by another participant, in order to be able to fix our perception of reality when it goes wrong. We cannot ever assume that, once we have interpreted a participant's utterance, our interpretation is infallibly correct, nor indeed can we expect our hearers automatically to make the appropriate deduction of our intended meaning. We must be able to backtrack to correct mistakes (which might then also have a knock-on effect on our interpretation of the conversation subsequent to the error).

The hypothesis that language is a co-ordinated activity is therefore in disagreement with the claim that participants in a conversation attain mutual knowledge or beliefs, and I claim that this distinction is important for the realistic modelling of conversation. What I propose is a unified approach to the two roles of speaker and hearer in a conversation, but that still allows for speaker meaning and hearer understanding to come apart. In order to introduce this new model, in Section 8.2 I look at the general properties of co-ordinated activities, to see how far we can say that language behaves in a similar way, and to apply observations about other activities to the interpretation of language.

8.2 Language as a Co-ordinated Activity

An activity is defined as a sequence of related actions, which perform (at least) one overarching goal; if you are engaged in an activity, then you can be said to be in a state of action. A co-ordinated activity is characterised by there being two or more agents that take turns to carry out the individual actions that contribute towards the achievement of the activity. Even though their individual goals may be different, agents will still need to co-ordinate with each other in order to perform the activity. For example, in a game of chess, agents are in competition with each other, but in order to achieve the activity, which is playing chess, they still have to obey the rules of the game (I shall come back to this example shortly). What evidence can we find to indicate that language is a form of co-ordinated activity?

When you are describing a conversation that you have had with someone else to a third party, you are more likely to describe the actions you were performing in speaking and the reasons for engaging in the conversation, than to report the actual words you used in doing so; the action is more important than the detail. So, for example, if you ring someone up, and I ask later why you did so, you are most likely to say "I needed John's phone number" or "To ask Jackie out to the pub tonight", because the actual language you utilised is of secondary importance to the actions you performed by your communication. Language is, in other words, a means to an end (which is the performing of an action or actions), the end being more important than the means. This makes sense when one thinks of the variety of different ways there are of performing the same act in speech. So, when reporting any speech activity, we are highly unlikely to utilise the original words; we reconstruct and paraphrase the content from our memory of the acts that were performed, probably because this is a much more efficient method of storing information. Speech acts are the means by which we compress language. Like other types of activity, each successive action that forms a part of it, adds and builds on what went before.

Language often has more than one layer of activity; in many different types of conversation, there can be more than one domain of action. For example, if I tell a story, it is important that my hearers recognise that I am in fact telling a story on the underlying layer of activity, but also follow the action that goes on within the story. This mechanism of layering our understanding is likely to be related to the way we swap between different topics, or nest them within our speech. The hearer will patiently wait to find out how a story, or interruption, relates to the conversational structure, just as he will participate in a nested conversation in the expectation that the current exchange will contribute to the goal of the higher level conversation segment from which it originates. It is interesting, with relation to the discussion of this section, to note that we often tell stories as a sequence of actions and explanations of motivation. (I shall briefly revisit stories and storytelling as a particular characteristic of conversation in Chapter 10.)

If discourse is a type of co-ordinated activity, then this provides the motivation for studying and defining some of the general characteristics of other types of co-ordinated activity. In fact, we had better be able to define the general properties of co-ordinated actions in order to see if they fit in with the general properties of acts of communication. In Table 8.3, we can see some variations in the types of co-ordinated activity that we carry out in our day-to-day lives.

DIMENSION OF VARIATION	*FROM*	*TO*
Scripted vs. unscripted	marriage ceremony	chance encounter
Formal vs. informal	city council meeting	gossip session
Verbal vs. non-verbal	telephone call	football game
Co-operative vs. competitive	business transaction[7]	tennis match
Egalitarian vs. autocratic	making acquaintance	class lecture

Table 8.3 Dimensions within co-ordinated activities.

The first three 'dimensions' are taken from Levinson (1992), and the last two are Clark's (1996). There are of course many other factors that will affect the structure and vocabulary of an activity as well, so potentially the number of activity types is vast. However, these are a good starting point and cover the basic types.

According to this list of criteria, general conversation would be classified as: unscripted, informal, verbal, co-operative and, as far as possible, egalitarian. As such, conversation seems to be one of the least constrained forms of activity in terms of a fixed framework of actions to be followed. This is unlike the majority of other activities, which have prescribed actions and rules, even a prescribed order for those actions (think about for example the order of events in a marriage ceremony, a tennis match, or buying an item from a shop[8]). On the surface, general conversation does not appear to have these types of restrictions; but, as I hope to demonstrate, this is not entirely true.

8.2.1 Tracing the State of (a Co-ordinated) Activity

One of the features that all activities have in common is that, at any given time, they are in a particular 'state of activity' and one can 'trace' the moves that led to the current state of activity. An example from the real world would show this property of activity in sharper relief. For

---

[7] This is not a particularly good example of a co-operative activity; in business transactions, typically the participants will have different aims according to their own interests and will have to negotiate with each other to come to an agreement. However, as this list is taken from other sources, I shall here leave the example as it is.

[8] Interestingly, many human activities involve the use of speech in their performance.

instance, as mentioned previously, the game of chess can be seen as a kind of co-ordinated activity, because though players' individual aims may be in conflict (they will each wish to win the game), the higher goal, that of having a game of chess at all, has to be co-ordinated and rule-governed. I will return to this point later. At the end of each participant's turn, the current state of the board can be described by a list of individual moves that led to the position of the moment.

To illustrate this point more clearly, I use the famous game between an IBM computer program called Deep Blue and a human grandmaster, the reigning world chess champion, Garry Kasparov[9]. The moves are given in one of the standard types of chess annotation shorthand in Table 8.4 (for anyone unfamiliar with this notation, the moves are shown graphically later, in Table 8.6).

1.	e4	c6	6.	Bd3	e6	11.	Bf4	b5	16.	Qd3	Bc6
2.	d4	d5	7.	N1f3	h6	12.	a4	Bb7	17.	Bf5	exf5
3.	Nc3	dxe4	8.	Nxe6	Qe7	13.	Re1	Nd5	18.	Rxe7	Bxe7
4.	Nxe4	Nd7	9.	0-0	fxe6	14.	Bg3	Kc8	19.	c4	1-0
5.	Ng5	Ngf6	10.	Bg6+	Kd8	15.	axb5	cxb5			

Table 8.4 Deep Blue (White) – Kasparov (Black), Game 6, 11th May 1997.

The two players begin the game with the chessboard in its initial state SA_0 (where SA stands for 'state of activity'). White plays first by making the move M_1, shifting the king's pawn two squares forwards. By the move M_1, white changes the state of activity from SA_0 to SA_1. What white does is *increment* the state of activity:

$$SA_0 + M_1 = SA_1$$

Now it is black's turn. He moves his queen's bishop pawn one square forward. This time, it is the state of activity SA_1 (not SA_0) that is incremented by this second move, M_2, to produce a new state of activity SA_2. Then white moves its queen's pawn forward two squares, M_3, which increments SA_2 to produce SA_3, and so on. A game of chess can be seen as the cumulative effects of moves added together to produce the current state of activity, and the formula above can be generalised to reflect this in the following way:

$$SA_{i-1} + M_i = SA_i \qquad (i \geq 1)$$

---

9 This game is a significant milestone in the history of the development of artificial intelligence techniques, because it is the sixth and last of the tournament in which, for the first time, a computer beat a human grandmaster at chess (albeit under unusual conditions).

This can be read as: each successive move M_i is added to the previous state of activity SA_{i-1}, to produce a new state of activity, SA_i. Table 8.5 shows how the sequence of moves described by the formula are performed. Note that in a game of chess, it is always the case that each move will signal a change of player, from white to black or vice versa, and that move M_1 (and, in fact, M_i when *i* is odd) will always be made by the white player. In other words, the order of play, and whose turn it is, is strictly delimited.

TIME	MOVE	STATE OF ACTIVITY
0	Open game	SA_0
1	M_1	SA_1
2	M_2	SA_2
3	M_3	SA_3
…	…	…
n	M_n	SA_n

Table 8.5 State-trace, for a game of chess for example.

Given the initial state SA_0, you can represent the course of the game by these two sequences:

States of activity: $SA_1, SA_2, SA_3 \ldots, SA_n$
Trace of activity: $M_1, M_2, M_3, \ldots, M_n$

Moreover, given the initial state SA_0, you would only need one of these sequences to be able to derive the other.

The state-trace representation of chess is demonstrated visually in Table 8.6, which is obtained by combining the game in Table 8.4 with the model outline defined in Table 8.5, replacing M_i with the chess shorthand annotation signifying the moves and SA_i with pictures of the chessboard positions after each move M_i.

	White Move & State of Activity		Black Move & State of Activity			White Move & State of Activity		Black Move & State of Activity	
0	Open game				**10**	Bg6+		Kd8	
1	e4		c6		**11**	Bf4		b5	
2	d4		d5		**12**	a4		Bb7	
3	Nc3		dxe4		**13**	Re1		Nd5	
4	Nxe4		Nd7		**14**	Bg3		Kc8	
5	Ng5		Ngf6		**15**	axb5		cxb5	
6	Bd3		e6		**16**	Qd3		Bc6	
7	N1f3		h6		**17**	Bf5		exf5	
8	Nxe6		Qe7		**18**	Rxe7		Bxe7	
9	0-0		fxe6		**19**	c4			1-0 **BLACK RESIGNS**

Table 8.6 State-trace for the game of chess shown in Table 8.4.

Note that in Table 8.6 (chiefly for the sake of getting all the moves on one page), I have included two moves in each chess line, whereas in Table 8.5 each move has its own event order index. One should actually read this differently (as shown in Table 8.7), because of course the two moves shown under each numbered line do not take place at the same time, and the order of play is important.

TIME	MOVE	STATE OF ACTIVITY
0	Open game	SA_0 = 
1(w)	$M_{1(w)}$ = e4	$SA_{1(w)}$ = 
1(b)	$M_{1(b)}$ = c6	$SA_{1(b)}$ = 
…	…	…
19(w)	$M_{19(w)}$ = c4	$SA_{19(w)}$ = 
19(b)	$M_{19(b)}$ = Resign	$SA_{19(b)} = SA_{19(w)}$

Table 8.7 Interpretation of Table 8.6[10].

This formal representation of chess is intuitively a very satisfying and neat description of the procedure of the game. We should note however that chess is an activity that is fairly rigidly structured; it might be less easy to define an activity such as a game of football, or a class lecture, in this way using the same straightforward notation. However, the point I would like to draw out here is that language, and more specifically the language used in general conversation, has a great deal in common as an activity with a game of chess, much more so than with a more physical and less structured activity such as football, as I will show in Section 8.2.2.

8.2.2 Comparing the Activities of Chess and Conversation

In this section, I shall look further at how far one can take the parallel between chess and conversation: what are the similarities, and what the differences? Can we compare the two activities in the same manner, and can we come up with a corresponding representation for language use, as we have done already for the co-ordinated activity of playing a game of chess?

---

[10] (w) stands for 'white' and (b) stands for 'black'. How the time index is labelled is not the important issue here. What is important is that the moves are ordered and the order of play matters crucially.

I have taken this method of comparing chess with language from Clark (1996)[11], but he compares the moves (the physical act of moving the pieces in chess) to the use of utterances in order to perform goals in a transaction, such as buying some item in a shop for instance; the state of activity (the current state of the chessboard in a game of chess), he compares to the state of the physical surroundings of the transaction, such as, which participant has the item to be bought, whether or not the item has been paid for yet, etc. It is interesting to note that there are often conventional patterns of behaviour that are expected and dictated by the situation: a sort of frame or script for the conversation to follow. In such situations there are already many expectations about acceptable turns in dialogue (for example, if you want to ask for timetable information at a railway station, or to open a bank account, or to ask the way somewhere). There is always a certain amount of scene-setting in any interaction, which then restricts our behaviour; to a certain extent this goes on in general conversation too.

However, I am not interested here in the application of the analogy of a game of chess to such a high-level, behavioural framework of activity – this in my opinion has more to do with the processing of high-level goals and inferences, as well as the planning of moves in order to fulfil a participant's overall plan. What interests me is the possibility of comparing a game of chess with the speech acts one performs in order to change our internal, mental state of activity in a conversation. This is because, in a non-transactional setting, it is not the state of the physical environment that, in the main, guides our language production and recognition.

So, with this in mind, let us informally compare some of the features of chess and conversation. In chess, players are not free to make any move they like, but are constrained by the rules that govern the way the pieces are allowed to move and that constitute playing the game of chess. Not only this, but the current configuration of pieces on the board (the state of activity) also restricts the choice of next move, i.e. the existence of a piece to move in the originating position and whether or not the intended destination position for a particular piece is clear of one's own pieces, etc. These two things (the rules about the way a piece should move, and the current chessboard position) give us a list of 'possible' moves in the context. For example, Figure 8.3 shows all the possible moves for an initiating player (which in chess as I have mentioned, is always the white player).

---

[11] Though the idea itself goes back at least as far as Saussure (1916).

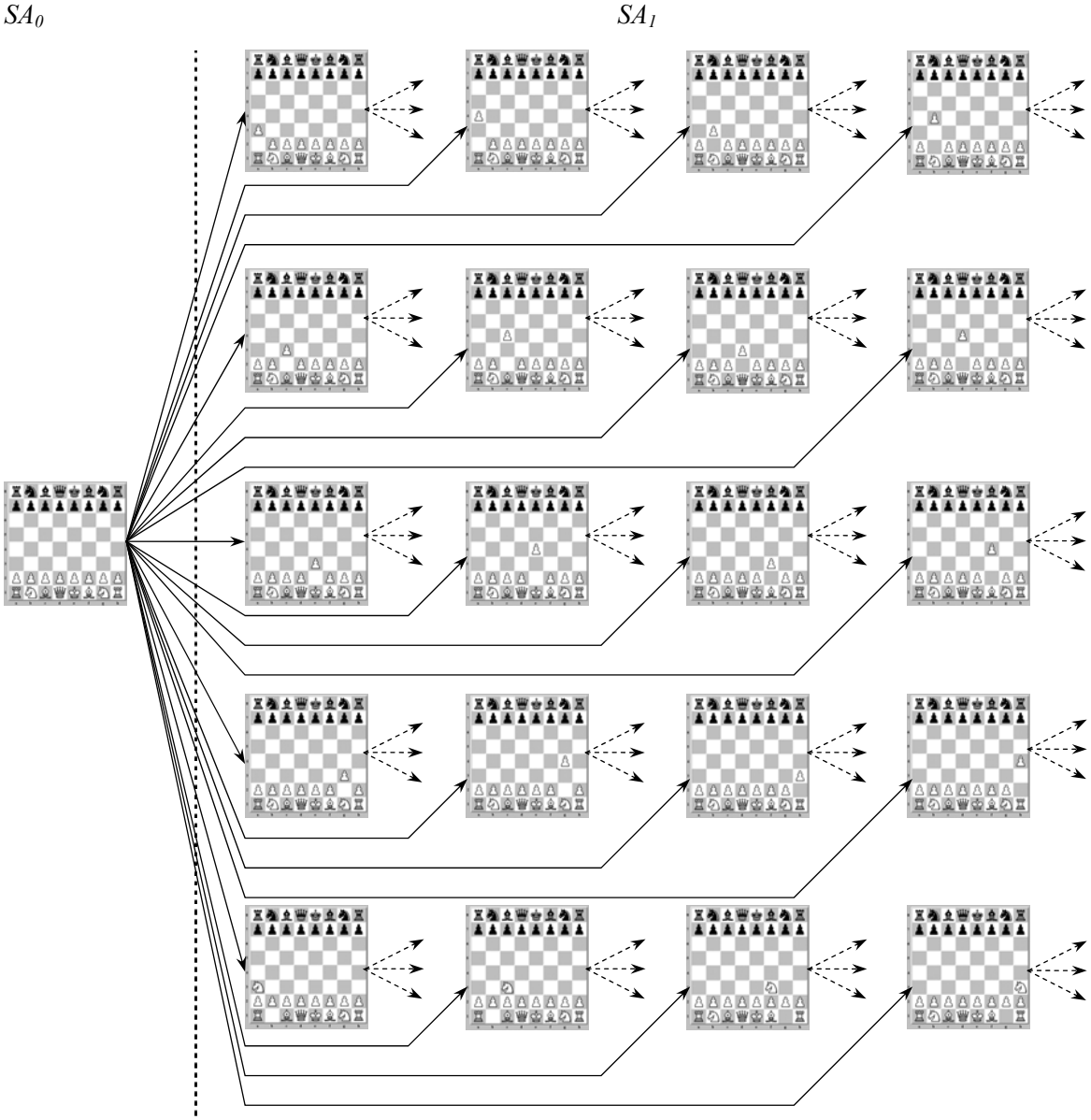


Figure 8.3 All possible choices of first move in a game of chess[12].

One can tell from Figure 8.3 that in chess the search space for investigating potential future moves soon becomes unreasonably large. There are twenty possible initiating moves, multiplied by twenty possible responding moves, and that is only after two turns in the game. This is one of the reasons why, for years, chess was considered an activity that was computationally intractable. However, from any one state of activity, there is only a constrained set of moves available to a player at any one time, and of this set many moves are not 'sensible' ones. We choose from this list of moves in accordance with our game strategy. In conversation too, while the choice of topic and content may not be finite, I believe that the choice of possible functions for an utterance are. In order to facilitate coherent and efficient speech exchanges,

---

[12] Of course resignation is also a possible, if highly unlikely, first move!

speech acts are constrained by the current context, and respond to a prior act, unless initiating a new 'game'.

In practice, some moves are conventionally more likely to be chosen than others. Knowledge of openings and closings saves processing time, because one has learnt the games that cannot be played out from the current position. The more one plays chess (and especially the more one plays against the same opponent) the more skilled one gets at guessing what the other player's strategy is likely to be.

Moves are very much related to one another although it may not appear so at first. While responses are not as obviously linked as they are in a conversation, if I attack one of your pieces, it would be wise of you to respond by protecting it, or by counter-attacking one of my pieces.

At the level of chess played by grandmasters and world champions like Kasparov, the relationship between different moves may be harder to spot, as they make defensive moves in anticipation of possible attacks, so moves seem less connected. People perform similar moves in conversation when they know the person with whom they are interacting well, and can predict some of their patterns of behaviour (either from prior interactions or the immediate physical context). This explains utterances such as:

> 'Before you ask me, I haven't had time yet to fill in the forms.'

But in general, when not actually initiating a conversational 'game' themselves, people respond to the most immediate move in the conversational context. Referential immediacy helps maintain threads and is less computationally expensive, which is important when the aim is effective, quick communication (Sperber and Wilson 1986). This is not to say that nested topics do not occur, for plainly they do (see Conversation 1.1 in Chapter 1 for an example), but we tend to treat them as separate games – we suspend activity in one while we make moves in the other, before returning to the previous game. In effect, we build a stack of open topics – this is much like the ideas expressed by Reichman (1985) and Grosz and Sidner (1986), as well as other kinds of finite state networks, nested finite state networks, recursive and augmented transition networks. In a way chess is also like this in that, although the overall aim is to win, this is achieved (or not, as the case may be) in stages, by localised battles for positions on the board.

At any given time point in the game, the current state of the board SA_i only gives you a summary, or the functional 'gist' or record, of the game that has been played so far – you cannot with any degree of certainty reconstruct the exact order of play from any one SA_i, as there may well be several 'paths' to the same state of play. The state of activity informs the decision making for the kind of move you should make next; a player need not remember the exact

sequence of moves if he can interpret the state of the board in order to tell how he is doing. So the physical representation provides a shortcut or memory aid for the players. It is the ability to assess your current position and plan the moves that will help you meet your goal (which in chess is to win by bringing the opponent's king under an inescapable attack or 'checkmate') that makes a good player. This kind of representation is key to the approach I have taken in this dissertation to the identification of speech acts. As I hope to demonstrate, this is, in actual fact, the fundamental idea behind the theory of context change (with a few minor differences). This is because, in the approach I have adopted, I have posited an internal representation that is analogous to the physical representation of the chessboard, so that participants in a conversation work out what speech act to perform next from this, rather than the entire utterance history (see Chapter 9).

Note that this representation of the game has nothing whatever to do with the various motivations that might exist behind each move that is made. There is no accounting for the fact that the moves are not at all random, but generally the first step in a plan to gain an advantage that leads a chess player closer to his ultimate goal, which is of course to win the game.

A player may further interpret a move, for example, 'attacking the knight', 'pinning the rook', or 'defending the queen'. These descriptions of purpose or intent can themselves be grouped into larger sequences that describe aims to weaken the opponent's position (because after all, chess is a competitive activity; while this might be compared to an argument or discussion, it is unlike general speech, which is normally at least negotiative). We might also note that some moves are seen as 'mistakes', or 'brave', or 'surprising', which adds a further level of interpretation to the model. (Might this be seen as a parallel with perlocutionary effects?) In the same way that one might produce a précis of a conversation after the event as a list of actions (rather than the exact words used), one might also produce an 'outline' of the play in a game of chess from the descriptions of purpose inferred by the player, e.g.: 'Opened by exchanging pawns; black brought queen out; exchanged pawns again and black queen put white in check', and so on. This abstracts away from the individual moves, but tracks the important actions, much like telling a story does.

Conversation, while having many parallels to an activity such as chess, also has a number of features that do not match. Part of the skill of playing chess is to be able to assess what a move 'means' in terms of your opponent's plan of attack. Far from wishing your co-participant in a game of chess to perceive your plan, it is in your favour that he be unable to do so, and vice versa for him. While a chess player hopes to keep his plans a secret from his co-participant, a speaker aims at (at least apparent) transparency between himself and the other participants about his goals and objectives. It is in a speaker's interests to achieve an understanding with his hearer – communication is essentially collaborative and rhetorical (in the sense of Leech 1983).

Even if a speaker is being misleading, he still means his hearer to understand (as sincere) his deceitful act, and the mechanisms for inferring speaker meaning are geared up for honest, co-operative behaviour. So, chess is a competitive activity, whereas, for the most part, communication tends to be collaborative and co-operative in nature.

The scope of possible moves in chess is much wider than those in conversation[13], and each player makes exactly one move each in regular turn. There is no question of usurping your fellow player's go, because the rule that a player should have exactly one move per turn is constitutive of the game of chess. To move three times would be deemed cheating! Also, each chess move is discrete in time, whereas often utterances overlap or are produced simultaneously. Acknowledgements for example regularly overlap with what is being acknowledged. In other words, the rules of chess are such that one cannot jump out of the current 'game'; chess is a static and relatively closed system, whereas conversation is a dynamic and open one.

Chess, as I have noted at the start of this section, has a physical representation of the state of activity – there is therefore rarely ever any confusion between the different players about which moves are legal next, precisely because there is an unambiguous external representation of the context for the following move. Conversation however relies on an internal representation of the conversational context; the nearest analogy for chess would be if both players were blindfolded and had to tell each other verbally which moves they make. Or, vice versa, an external representation of a discourse might be rather like taking notes (the minutes of a meeting for example) or producing an actual transcription or recording of the conversation that can be consulted by the participants before each utterance is produced. Note that, as I discussed in Section 8.1, even when a recording has been made, one might still be unable to reproduce the intended meaning of the conversation accurately. On a day-to-day basis though, this facility is simply unavailable to participants in conversation. It is for this reason therefore that I argue that we must keep some kind of internal model of the state of activity of our interactions.

I would suggest that there are few people who have the capacity to keep a model of the state of activity of something as complex as a game of chess in their heads for any length of time (I certainly could not play chess without an external representation – the chessboard and pieces – to help me). Although there are people with the memory and mental capacity to do so, I would suggest this is the rarity rather than the rule. We might conclude therefore that the structure of

---

[13] I am of course here talking in terms of function, not topic. To compare, while there are 20 ways of opening a chess game, I shall try to show in Chapter 9 that, in terms of function, there are only really 3 or 4 ways to initiate a conversation (if we discount for the moment such things as greetings, etc.). The conversational 'games' we play are therefore much simpler in this sense.

conversation must be considerably simpler to allow for our limited memory capacity. So, if conversation is a type of game, or has the same structures as a game, the games must be short and the rules relatively simple (cf. Footnote 13).

As I have shown, the similarities between chess and discourse are quite striking; discourse can even be represented using the same kind of structural format as for the moves in chess given in Section 8.2.1, as demonstrated in Table 8.8.

TIME	**UTTERANCE EVENT**	**CONTEXT**
0	Open conversation	C_0
1	UE_1	C_1
2	UE_2	C_2
3	UE_3	C_3
…	…	…
n	UE_n	C_n

Table 8.8 State-trace for a conversation.

This representation does not seem to work as well as it does for a game of chess (as shown in Table 8.5), because there seems to be at least one step missing. In chess, it is the physical act of moving a piece through space that changes the state of activity, whereas in speaking it is the speech act interpretation of the utterance event, which then maps the event onto the new context. So with this added layer of interpretation, we would amend the formula given in Section 8.2.1 to:

$$C_{i-1} + UE_i \rightarrow C_i \text{ by } SAct_i$$
$$\rightarrow C_i' \text{ by } SAct_i'$$
$$\rightarrow C_i'' \text{ by } SAct_i'' \ldots$$

Or: $C_{i-1} + SAct_i(UE_i) \rightarrow C_i \qquad (i \geq 1)$

This reads: the previous context (C_{i-1}) is incremented by the current utterance event (UE_i) to produce a new context (C_i) by the interpretation of the utterance's speech act ($SAct_i$); or else some other new context (C_i') is produced by some other speech act interpretation ($SAct_i'$), etc.

In other words, it is not enough to look at just the utterance event, you have to know the function represented by that event in order to be able to update the context correctly. It depends on your interpretation of the act performed how the context gets updated (what effect an utterance has on the context/state of activity). Because the process of ascribing an interpretation is not completely transparent, this explains why people can sometimes understand different things from the same speech exchange.

Clark concludes from comparing an activity such as a game of chess to conversation that you cannot distinguish between the characteristics of discourse and any other 'co-ordinated activity'. This is a very large claim to make and, although I agree in principle that one can approach language as one can any other type of activity, there are also certain features that are unique to discourse (such as the non-visual nature of conversation) that must also be taken into account.

8.2.3 Defining the Context of Language

There are at least four different types of contextual information that are required in order to interpret a speaker's utterance as he intends:

(1) *Initial (assumed) common ground*: The set of background 'facts', assumptions and beliefs presupposed by the participants (including knowledge of the physical setting), although these are not necessarily correct or the same for all participants.

(2) *Public (utterance) events so far*: A list of public events (which, in the case of conversation, equate to utterances, UE_i) that have occurred in order to produce the current state of activity.

(3) *Current state of the (co-ordinated) activity or conversational context*: That which the participants suppose is the state of the conversation at a particular time (the context C_i).

(4) *Translation rules that define the effect of the events on the context*: This is, in effect, the discourse model that transforms an utterance into a speech act (the $SAct_i$) that changes the state of activity. In order to be able to communicate effectively, this model must be common to all speakers (of the same language).

My research is not concerned with (1) above, but with (4), how we recognise the functions that map the occurrence of an utterance event UE_i onto the changed context C_i. The question is, is it really feasible to ignore (1) totally and get any reasonable and meaningful results? Suppose that (1) contributes crucially to the recognition of the function of an utterance? Can one ever divorce the spoken context from the factual and physical one?

These issues are not straightforwardly answerable. There seems to me little doubt that the ultimate interpretation of an utterance's meaning will almost certainly require recourse to world knowledge in the majority of cases. But here I must remind the reader that it is not the inferential mechanism that I am in the process of elucidating, but the structural mechanism. I wish to see how far it is possible to abstract away from the content of a sequence of utterances to look at the function: not so much the 'what' and the 'why', but the 'how'.

Interestingly, as I have mentioned previously, Clark does not seem to distinguish the two different types of context (knowledge of the world and knowledge of the conversation) but conflates them into one – I believe mistakenly so. He suggests that the physical scene in which

the conversation takes place is the equivalent of the physical representation of the state of play on a chessboard. This seems to misunderstand essentially the nature of action in spoken transaction, because plainly this is not the correct parallel to draw. The state of play of a conversation has no external representation, and I would argue strongly that this is why it is subject to error.

Although one might criticise approaches to conversation that do not take background and physical context into account (mine included), when we are seeking to represent contextual information computationally, it is currently impossible to encompass comprehensive world knowledge, even only that knowledge typically at the command of a single human being. Some kind of compromise must be made, but not at the expense of being able to expand the model to include an account of background as necessary.

8.3 Summary

In this chapter I have tried to establish the following two hypotheses (in order to set up the presentation of a formal model of speech acts in Chapter 9):

(1) *Language use always involves speaker's meaning and addressee's understanding*: When a participant says something, he means something, some message, that he intends the hearer to understand (unless of course he is schizophrenic). This is central to language use – a lot of effort goes into making sure that we are communicating effectively with each other when we speak. However, we can never guarantee that our utterances will be perfectly understood. We can only reach a level of adequate understanding (as discussed in Chapter 2).

    There is a basic assumption that the way two participants perceive the interaction will lead to adequately similar models of context; communication proceeds on this assumption until it becomes obvious that this is not the case, when a repair sequence will be initiated.

(2) *Language use is a species of co-ordinated activity*: Most language use requires a minimum of two agents – communication is not just individuals performing 'autonomous' actions, but a co-ordinated collaboration between participants towards the working whole. It is like a jazz band improvising a piece of music together – all parts of the band must be responsive to each other and take turns to lead, otherwise the end result would be a cacophony[14].

---

[14] Some might say that jazz is a cacophony – however, this is a serious point. Jazz is a 'messy', improvised kind of (musical) activity, but so is conversation. What prevents it being chaotic is an underlying co-operative structure.

Although these observations may seem commonsensical, traditional approaches to speech act theory have so far failed to account for the part that role-switching in conversation plays in the identification of an utterance's function. Seeing language as a type of game is by no means a new idea, but if we accept that communication is a form of structured co-ordinated activity, then it might be possible to characterise exhaustively the functional 'moves' in this language game. The questions at the heart of the theory of language as action are: what comprises a comprehensive set of speech acts, how do we identify speech acts from an utterance, and how can we define the changes that occur to the context or state of activity after the performance of a speech act so that we can be sure that we have captured the full functionality of language?

As we have seen in this chapter, in an activity such as the game of chess, it is possible to define the context of the next move precisely, and because of this, all the possible choices of following move as well, as there is only a finite set of moves that one can make at any one point in the game. Of course, in practice one could never generate all possible games of chess, because some games of chess have the potential to be infinitely long. Humans generally recognise when a game has gone into an infinite loop, realise that neither player can achieve the goal of winning outright, so choose the next best option of drawing the game to get out of the stalemate position. In fact, in a conversational 'game', the same sort of phenomenon can occur too; e.g., when participants realise that they can never come to an agreement concerning a particular proposition, they then have to agree to disagree, or find some compromise between the two positions that is satisfactory to all parties. This is effectively the same as drawing the game.

So, the question is, can we define the context of language in such a way that we could list all possible next moves that participants might make in *any* possible situation? On the face of it, this seems impossible. Humans are not restricted in terms of the sort of thing they can talk about – the topics of possible discussion are endless, and the ways of formulating an utterance are various, etc. However, functionally, I will argue that language behaves in very much the same way as a game, and as such, it should in principle be possible to produce the necessary constraints so that a finite set of expected next moves could be generated. If I can show that this is indeed possible, then I think that I shall have gone some considerable way towards demonstrating how we compute the meaning of an utterance in a conversation, and how it might be done automatically. This is what I shall attempt to do in Chapter 9.

Chapter 9

Modelling Speech Acts in Context

> For all meanings, we know, depend on the key of interpretation.
> (*Daniel Deronda*, George Eliot)

> …lo que conoces
> es la tristeza
> de mi casa vista de afuera.
>
> …what you know
> is the sadness
> of my house seen from outside.
>
> (*Es tan poco* – It's so little, Mario Benedetti)

In this penultimate chapter, I shall present the theoretical model of speech act use in conversation that has been developed as a result of the observations I have made in my research and described so far. I shall exemplify how the model works by processing segments of the conversations that I have collected, and finally by analysing Conversation 1.1 from Chapter 1. This will highlight both the advantages and drawbacks of using this kind of approach in characterising the role of speech act interpretation in conversational discourse.

Before doing so however, it would be beneficial at this stage to clarify exactly what I mean by the term 'speech act' in this dissertation. Often in the literature the definition of a speech act is left vague and different types of phenomena are conflated under the term.

9.1 What Constitutes a Speech Act?

Defining the exact constitution of a speech act is not straightforward. An utterance can perform a number of different kinds of actions within a conversation, as has been consistently shown throughout this dissertation and in the literature that is surveyed within it. The multi-functional nature of utterances is neither a very controversial nor a particularly original notion. However, the problem that still remains to be addressed is that, as speech acts are abstract functional entities, how is it possible to be certain that one has defined them all? How can one be sure that one is dealing with the abstract functional entity at the right level? What, in short, is the make up of a speech act, and how does it relate to the structure of discourse? Searle (1979: 29) also notices this confusion:

> There are not, as Wittgenstein (on one possible interpretation) and many others have claimed, an infinite or indefinite number of language games or uses of language. Rather, the illusion of limitless uses of language is engendered by an enormous unclarity about what constitutes the criteria for delimiting one language game or use of language from another.

In this section I shall try to highlight and break down this problem to show how we might be able to classify and analyse speech acts comprehensively at a certain level of abstraction. I begin by explaining what I mean by the definition of speech acts at different 'levels of abstraction'.

9.1.1 Levels of Abstraction

Few researchers working with models of discourse using speech act theory make the distinction between different types and levels of speech acts that are performed by the production of an utterance in a speech situation. Whilst many admit that an utterance may perform more than one act at the same time, this is rarely explained in terms of how an utterance manipulates the context at different levels of abstraction. I am convinced that one of the reasons behind some of the confusion and dissention surrounding speech act theory today has been brought about because researchers often do not deal with phenomenological entities at the same level of analysis. I believe that what kind of thing a speech act really is has never been rigorously defined.

Bunt (see Section 6.2.3) distinguishes speech acts performing at different levels in his classification of different types of context, and in the split between the task-oriented and dialogue control functions of utterances. Traum and Hinkelman (1992) also identify different acts performing at different levels. Other schemes and systems for the recognition of speech acts in discourse (especially statistically based models) are currently hampered by the need to specify one and only one act per utterance and also by the lack of any recognition that an utterance may perform acts at different levels of abstraction at the same time[1]. I argue that it is the combination of these functions at different levels that allow an utterance to have overloaded meanings and enables extremely compact communication to take place.

This is not to say that the multi-functionality of utterances goes unrecognised. The problem has been identified especially by those who take an interest in speech act theory from a socio-interactive point of view (e.g. Halliday 1975, Labov and Fanshel 1977, and Schegloff 1987), and is expressed very succinctly by my namesake Schiffrin (1994: 86):

---

[1] For example, the mixing up of dialogue acts such as BACKCHANNEL/ACKNOWLEDGE and AGREEMENT/ACCEPT in the SWBD-DAMSL scheme by annotators can be explained and possibly avoided altogether by treating them as instances of each other at different levels of abstraction. In agreeing with some other participant, I am also acknowledging their previous utterance.

> …do we want to say that all of the many functions realized through a single utterance are speech acts? And do we then need to include all of these functions in the classificatory schema of speech acts assumed to be part of communicative competence?

This observation of Schiffrin's (1994) comes out of her discussion of speech act theory in an overview of various approaches to discourse analysis. Using an example of the different functional breakdowns one can obtain of the same utterance sequence, she relates her analysis to Searle's felicity conditions for the performance of certain speech acts in order to explain why and how an utterance can masquerade as more than one speech act at the same time. The following example conversation is taken from Schiffrin (1994: 85):

Utterances		*Sequence 1*	*2*	*3*
Henry:	(1) *Y'want a piece of candy?*	QUESTION	REQUEST	OFFER
Irene:	(2) *No.*	ANSWER	COMPLIANCE	REFUSAL
Zelda:	(3) *She's on a diet.*	{Expansion}	Account	Account
Irene:	(4) *I'm on a diet.* + story	{Expansion}	Account	Account

This conversation is analysed as three separate sequences following on from utterance (1), which might be interpreted as:

(a) A QUESTION that sets up the expectation of a 'Yes' or 'No' answer, which is then fulfilled by Irene's response 'No' in (2).

(b) A REQUEST for information that sets up the expectation of the provision of the necessary information.

(c) An OFFER of a piece of candy, which sets up the expectation of acceptance or rejection of the offer, which is again fulfilled in (2).

Schiffrin deduces these different prescriptive[2] functions (QUESTION, REQUEST and OFFER) by matching the conditions listed by Searle that must hold true in the current situation for the achievement of that function by a certain utterance. I shall not provide a full specification of the

---

[2] It could be further noted that though the interpretations presented here are all prescriptive in nature (they prescribe primarily an answer to the QUESTION, REQUEST and OFFER), the grammatical form of utterance itself is structurally declarative. If we were to take it literally, Henry would be asserting that Irene wants a piece of candy. Presumably intonation here removes this interpretation from consideration and indicates that this utterance is the contracted form of 'Do you want a piece of candy?'; the addition of a question mark in the transcription makes this clear, although we could probably have concluded this from semantic knowledge – Henry cannot know whether Irene wants a piece of candy because he has no direct access to her mental desires. For the interpretation of an utterance such as this one, context is vital. I shall return to this in Section 9.2.4.5.

rules for these acts here as Schiffrin covers these. I am a little sceptical of this approach to the recognition of an utterance's function, because I am not convinced that there is a huge difference between a QUESTION, a REQUEST and an OFFER; I will argue that one of the reasons that this utterance can be functionally overloaded is that the functions are structurally related. I will discuss this further later in this chapter. For now, I merely wish to demonstrate how it is commonly possible to functionally overload a single utterance. I believe that this approach can be significantly simplified and that the phenomenon illustrated by this example is at least partially caused by the conventional use of interrogatives in English to perform requests and offers. Again, I shall return to this latter point later.

For the present, I wish to draw out two salient observations from a conversation such as the one given in the example above. The first is to point out that nearly all utterances have an element of requesting about them in that, generally, someone says something in order to elicit some kind of response from their audience, even if it is only an acknowledgement of the utterance having been heard and (apparently) understood. The second is to note that in order to interpret (1) as an offer we must take the semantic interpretation of the verb 'want' into consideration. In this case it is necessary to know who the beneficiary of the action will be (see Section 10.1.4 for further discussion). In other words, while the first and second speech act sequences rely on syntax and intonation (an illocutionary force indicating device that, in this instance certainly, plays a vital part in the identification of a declarative utterance as a question) for the ascription of speech act function, the third sequence depends on conventional semantics, and is based on hard-wired recognition of purpose. In other words, the former interpretations are derived structurally (though not necessarily wholly so) while the latter is derived inferentially (even if conventionally so).

Smith and Holdcroft (1991) have attempted to deal with the overloading of function in speech by incorporating two different stages in their model of speech acts: the first stage appeals to any discourse expectations generated by earlier conversation; and the second stage appeals to conventional means of recognising a speech act. Reichman (1985) gives a formal description and an ATN (**A**ugmented **T**ransition **N**etwork) model of conversational moves, with reference to conventional methods for recognising the speech act of an utterance. She uses the analysis of linguistic markers such as pre-verbal 'please', modal auxiliaries, prosody, reference, clue phrases (such as 'Yes, but...' (sub-argument concession and counter argument), 'Yes, and...' (argument agreement and further support), 'No' and 'Yes' (disagreement/agreement), 'Because...' (support), etc.), other indicators of potential[3] illocutionary force. My work on a

---

[3] I use the word 'potential', because even these linguistic items are subject to contextual interpretation.

model of speech acts would have to act as a 'front end' to her model, by providing a first level of utterance interpretation on which the overall conversational structure can then be built. How feasible it is to abstract different levels of analysis in this way, and how much interplay there is between the identification of one level leading to the identification of another are matters for further consideration[4]. However, I hope to go some way towards answering this question by the end of this dissertation.

I shall not give a thorough account of such cue phrases, as various other researchers (Reichman 1985, Stenström 1994 for instance) have already dealt with these admirably. I will however consider the way some of these phrases work at appropriate points in the discussion in Chapter 10. For the purposes of describing the functioning of my model, I assume that this kind of processing of utterances has already taken place. There are many reasons why I have limited myself in this way. Quite apart from questions of scope (to include all of the above listed methods of speech act identification is work beyond the capabilities of any one theory at this present time), I also wished to keep the model as simple and uncluttered as possible, and yet make it powerful enough to demonstrate the potential of the theory. I am not so much interested in the forms of expression of speech act utterances, but rather in the overall interplay of speech acts in a conversation – a syntactic structure of discourse, so to speak. I recognise that one can only get a partial functional analysis of conversation by ignoring the form of an utterance in the interplay of speech acts, but in Chapter 10 I will discuss some of the ways in which the model could be extended to produce a richer functional analysis.

So, we have discussed how there may be ways of building related and co-dependent kinds speech acts from a basic type of function by the use of particular turns of phrase or verbs. However, there are other ways in which speech acts perform at different levels of abstraction; for example, when an utterance performs a spoken act that is interpreted, usually conventionally, in a very specific way. These types of acts are commonly associated with social functions, or with dialogue control (Bunt 1995). Some examples of these are:

(1) Self-introduction
(2) Other-introduction
(3) Greetings
(4) Leave-takings
(5) Apologies
(6) Expressions of gratitude, etc.

---

[4] There are overtones here of the more general debate over cognitive architectures exemplified by the modularity of Fodor (1983) as opposed to the connectionism of Rumelhart et al. (1986).

Some of these so-called 'speech acts' do not in fact commit us to anything in their performance; they are content free. E.g.:

'Hello' ∅ **GREET**

Others have two distinct types of speech function:

'My name is Mandy' } ASSERT ('Speaker's name is Mandy') **SELF-INTRODUCE**
'I'm/am Mandy' }

There is something significantly different about these acts. They are evidently performing more than one role in the conversation. The first act (ASSERT in the example above) is the basic interpretation of function, and the other is a more complex interpretation based on the recognition that the basic function can be used as an instance of a particular behaviour (here it is SELF-INTRODUCE). This latter type of act, which is built on the former and represents a social formalism, is not the kind of phenomenon that I am principally interested in elucidating in this dissertation.

In this section, I have tried to show that there is a problem with the label 'speech act' in that it has arguably been used as such a broad a term that it virtually encompasses any kind of act that is accomplished through spoken means at any level of analysis, be it syntactic, semantic or pragmatic. I have considered using some other label for the type of phenomenon I am concerned with in this dissertation (such as 'context-changing device' for instance) in order to differentiate what I mean, but I have instead decided to appropriate the term 'speech act' for my own use, to mean specifically the kind of basic structural function that maps the state of the conversational activity prior to the current utterance, onto the state of the conversational activity afterwards, as discussed in Section 8.2.

9.1.2 Commitment vs. Belief

What I am proposing is a model of speech acts that takes the conversational context into account, and works in a purely structurally pragmatic way. At this stage, I would like to emphasise that, unlike many researchers in this field, I am not counting beliefs and background knowledge as part of the context of conversation. This is not to say that I do not recognise the importance of these in identifying the intention of any utterance, but I would like to see how far one can proceed without recourse to anything other than the conversation itself. Part of the justification for restricting the work in this way (apart from considerations of manageability) is that there is no mechanism in human communication by which normal human beings can access what each other believes - that is to say, the way people judge each other's beliefs is typically by what they say (and, presumably, do).

So each utterance of a speaker can be analysed as some sort of **commitment** to the propositional content – which defines a speaker's attitude towards the context depending upon the illocutionary force of the utterance. Note that this idea of commitment has nothing to do with the truth or falsity of the content, à la Austin (1962). The idea is that if I say that *P*, then I am committed to *P* being the case, regardless of the actual truth-value of *P*.

The **context** of a conversation is sets of propositions representing the various commitments of each participant. Speech acts are the means by which the context is altered. This is known as the **context change theory of speech acts** (Gazdar 1981, Levinson 1983 and Bunt 1994) and is in essence fairly simple. Speech acts are assumed to be types of function that map one context onto another. If this is so, then there must be (set theoretic[5]) operations associated with every speech act that will determine the effect of an utterance on a set of propositions (or commitments) – i.e. the current context of the conversation. When someone utters something, the conversational context of the participants in a conversation will be changed accordingly. So, as in the above example, if I say that *P*, then the fact that I am committed to *P* is added to the context.

Most speech acts add some propositional content to the context. Levinson (1983: 277) lists three of these as follows:

> (i) An *assertion* that *p* is a function from a context where the speaker *S* is not committed to *p* (and perhaps, on a strong theory of assertion, where *H* the addressee does not know that *p*) into a context in which *S* is committed to the justified true belief that *p* (and, on the strong version, into one in which *H* does know that *p*).
>
> (ii) A *promise* that *p* is a function from a context where *S* is not committed to bringing about the state of affairs described in *p*, into one in which *S* is so committed.
>
> (iii) An *order* that *p* is a function from a context in which *H* is not required by *S* to bring about the state of affairs described by *p*, into one in which *H* is so required.

He gives the following example of a speech act that would have the effect of removing some proposition from the context:

> (iv) A *permission* that (or for) *p* is a function from a context in which that state of affairs described by *p* is prohibited, into one in which that state of affairs is not prohibited.

---

5 By 'set theoretic', I mean that each utterance places, or replaces, a commitment to the proposition in the current 'context space'. I shall be defining what I mean by 'context space' and explaining this process of updating later in the chapter.

I would actually argue that although *permission* might remove a previous commitment to the state of affairs represented by the prohibition of *p*, it is actually subsumed by the new state of affairs in which *p* is allowed; the context is not emptied, it is modified.

This approach is appealing on many levels, and captures the intuition that, for example, it makes no sense to permit something that has not been prohibited, or agree with something that someone has not previously asserted. Hamblin (1970, 1971) was one of the first to introduce the idea that contexts could be seen as sets of commitment stores (Hamblin 1970), or **commitment slates** (Hamblin 1971; see also Wallis 1994, Schiffrin 1995). A commitment slate is a set of propositions representing the commitments of a particular participant in a conversation. Commitment slates equate to what I have been calling the context of the conversation. Each participant keeps one of these slates of commitments, or contexts, and updates it accordingly after every utterance. (Every utterance concerning the current proposition under discussion subsumes any previous commitment we might have had.)

9.1.3 Direction of Fit

The speaker is committed to the utterance content, and this commitment is recorded on the context in a certain way, depending upon the speech act identified from the context. Speech acts may either express commitments to the way the world is already, or commitments to the way the speaker (or some other agent) wishes the world to be changed; this corresponds to Searle's idea of direction of fit (as I have described in Section 3.2; see also Humberstone 1992).

There have been some criticisms of the 'direction of fit' analysis of language. The main argument put forward against it (by Rae 1989, for example), is that this approach would reduce language to two dimensions. This argument states that much of the richness and expressibility of language would be lost using such a method of classification. However, my own inclination would be to claim that far from limiting language, direction of fit is key to understanding much of its purpose[6]. Admittedly, it seems a fairly simple way of analysing utterances, but simplicity is not a weakness, unless it leads to a lack of comprehensiveness. The only obviously serious omission in this part of Searle's theory is that there is no account of how questions (interrogatives) would be accommodated within the direction of fit hypothesis. However, I will try to show in Section 9.2.4 that they can indeed be incorporated into this framework, by treating them as a special case of world-to-word fit illocutionary acts.

---

[6] Note further that Rae was trying to justify his reasons for disqualifying speech act theory in general from part in his work on the role of *explanation* in conversation, and therefore in conversational analysis as a whole.

Searle gives four directions of fit (as described previously in Section 3.2): Null ∅, words-to-world↓, world-to-words↑, and double↕. The double↕ direction of fit accounts for *declarations* in Searle's theory – this fit accounts for those speech acts that, when performed felicitously, change the state of the world by means of their utterance (such as 'I now pronounce you man and wife' in a Christian marriage ceremony)[7]. This direction of fit is completely different in nature from other categories: all speech acts of this type are *per se* declarations of some kind that are dependent for their felicitous execution on some extra-linguistic institution. They bear very little resemblance to each other, and there seems to be no unifying feature to relate them (except perhaps that many may only be performed conventionally within the confines of some institutional ceremony). This would suggest that there is a good case to be made for treating declarations as totally distinct from the other more usual speech act types. For example, declarations can never be performed indirectly. When some object is consecrated, or a witness is sworn in, the act is carried out according to some pre-specified linguistic formula. Often the words of the formula have to be repeated exactly for the declaration to succeed, so there is no question of any indirect interpretation. Smith (1991: 16) suggests renaming Searle's category of declarations as 'Formal Speech Acts', or (reverting to Austin's original name) 'Performatives'.

The null ∅ direction of fit encompasses Searle's category of *expressive* speech acts. Again there is some question as to whether this category should be considered on the same level as the two remaining categories. Like the double↕ direction of fit category, they seem to perform acts that are not properly to be considered at the same level as the more basic speech acts. Bunt (1995) identifies expressives with what he calls 'dialogue control acts'. Often the utterance of an expressive, such as 'I'm sorry for causing you any inconvenience', is the performance of an act according to some social rule of politeness. However, there is no doubt that to some extent a speaker is committing himself in some way to being sorry (even if he is not really sorry), and to the idea that his actions caused the hearer inconvenience. One could almost take it as a straight assertion, because a natural reply might be 'Oh, you didn't cause me inconvenience', although this itself might be a conventional polite response.

There is an argument for including this category in the study of conversation at a different level of abstraction, counting them as ritual exchanges performed according to some convention of politeness (see Goffman 1972). It is difficult to say exactly what effect the expression of a sentence such as 'I'm happy to see you again' would have on the context. Perhaps such an

---

[7] Utterances that are formed using explicit illocutionary verbs might also be said to be performing declarations of sorts. If the speaker says 'I assert that lemons are blue', then the contents of the speech act are made true in the moment of its expression (not that lemons are blue, but that the speaker has so asserted). I discussed this phenomenon in Chapter 4.

utterance would count as a straightforward assertion of the content '*S* [speaker] is happy to see *A* [addressee] again', and the speaker would be committed to the truth of this expression. However, neither the speaker nor the addressee might care whether such an expression is true or not: it may be interpreted as a welcome indicating phrase, performed according to social conventions of politeness, regardless of whether the speaker is being genuine or not. For this reason I intend to leave out expressive utterances from my model of speech acts, in the belief that such phrases could well short-circuit speech act interpretation.

9.1.4 Descriptive and Prescriptive Speech Acts

This leaves us with speech acts of one of two directions of fit – words-to-world↓, or world-to-words↑. This idea corresponds very well with a lot of work that has been carried out both in natural language processing on developing plans for generating helpful responses and planning speech acts (Cohen and Perrault 1979, Allen 1983), and in automated theorem proving and logic in planning (Fikes & Nilsson 1971). One cannot help but make the analogy between the two directions of fit, and 'facts' and 'goals' in logic and theorem proving. On the one hand a fact is declared, or described, and on the other some participant (the speaker in the case of 'offering') is prescribed to bring about some action to make some 'goal' proposition the case. Here however the analogy stops, because the list of so-called 'facts' need not actually contain any facts at all. I may have been lying (or, if we are being charitable, mistaken) when I said X. Even so I have still expressed a commitment to the truth of X – a hearer would be justified to complain or correct my assertion if it later turns out to be false, but I am still committed to the proposition that I have expressed. Therefore, to call this type of speech act a fact, or factive, would be highly misleading. So I will use the term **descriptive speech act** to refer to words-to-world↓ commitments, because they purport to describe the way the world is (as perceived by the speaker); and the term **prescriptive speech act** to refer to world-to-words↑ commitments, because they prescribe that the world should be changed to fit the words that express the proposition.

Descriptive speech acts correspond to Searle's *assertive* speech acts; prescriptive speech acts correspond to *directive* and *commissive*[8] speech acts. I will treat directive and commissive speech acts as in essence performing the same role; the difference is that in the case of directive

---

[8] The term *commissive* as I describe it here actually corresponds to what I shall be referring to as *self-commissive* speech acts, where the agent of the action is the speaker. There is also a strange, and as far as I am aware, completely overlooked set of speech acts, which I shall label *other-commissive*, where the agent of the action is a participant who is neither the speaker nor the addressee. I will discuss this shortly in section 9.2.1 and show how these speech acts might be accommodated in my model. However, I do not pretend to have fully covered this phenomenon.

speech acts, the agent of the action prescribed in the content is a specific addressee (someone other than the speaker), whereas in the case of commissive speech acts, the agent is always the speaker. There is a third type of world-to-words speech acts (not listed by Searle as a separate case): what I shall call *requestive* acts, which are performed in the interrogative mood. I shall discuss these last types of acts in Sections 9.2.4.

9.2 A Preliminary State-Trace Model of Conversational Context

Having discussed the constitution of the speech act as I have defined it, I am now in a position to relate the model of tracing the state of a conversational activity as developed in Chapter 8 to a more formal model of speech acts.

9.2.1 The Development of the Model

If we are to define a comprehensive method of identifying speech acts, then for every utterance, we must be able to define the context under which that particular utterance might be performing a particular act. The problem is that, like the case of the chess game, the number of possible options after each turn becomes combinatorially large. Added to this, for each participant in the conversation, the number of combinations explodes again. So, must we conclude that it is impossible to list all contexts exhaustively? If the answer to this question is 'Yes', then how can we rigorously define the conditions for the performance of each act?

The answer is to restrict the context in a number of ways (I shall come back to why and how this is possible shortly). The context does not encompass the combination of all prior moves, but records each participant's current position. By 'context' I mean the preceding conversation that is of immediate relevance to the current utterance; not the entire conversational history, but the previous speech acts in the sequence leading up to the performance of this one. These need not be immediately adjacent to each other; chronological contiguity is not necessary.

Perhaps instead of the word 'context' I should coin the phrase 'context space' in the same sense as Reichman (1985) uses it. She argues that utterances can be viewed as *conversational moves* performed in *context spaces*; a change of topic (performed by a certain conversational move) creates a new context space in the conversation. This can be visualised by using a computational example. Imagine opening an application in a window and performing several operations from within this application, possibly opening a number of other applications in the process. After the opening of a 'window' (which is symbolic of the opening of a topic of conversation), the 'window' is used until it becomes superseded by either a related 'window' or by a non-related 'window'. The original 'window' will either continue to exist in the

background (the context) in the expectation that it will be returned to later, or it will be closed altogether. I give a visual representation of this process in Figure 1.1, Chapter 1.

It turns out that listing the possible combinations for *descriptive* speech acts is trivially straightforward, because each act is restricted only by who has said what previously. In other words, each descriptive utterance only has one parameter that is of importance to the recognition of following speech acts: namely, who was the speaker of the utterance. The options are exhausted three levels of moves in, so a comprehensive list covering all context spaces is not difficult to compile. (This will become clear in the coming discussion – see Figure 9.1 and Section 9.2.2.)

The problems really begin when we look at *prescriptive* utterances. Prescriptives are difficult to list exhaustively because we are not just dealing with a proposition. Each prescription has at least three parameters: the speaker, the addressee and the agent of the prescription. The possibilities proliferate alarmingly after just a few moves, especially with the addition of other participants to the conversation.

To get a feel for the problem, let us look at just the number of options with three conversational participants. There are three possible prescription types[9] that the initiating speaker can choose to make (s = speaker, a = addressee and h = a member of the set of participants, such that $h \neq s$ and $h \neq a$):

DIRECT:	s says (to a) that the agent, a, will bring it about that X
SELF-COMMIT:	s says (to a) that the agent, s, will bring it about that X
OTHER-COMMIT:	s says (to a) that the agent, h, will bring it about that X

So far this seems relatively straightforward, with the exception of the oddness of the speech act that I have labelled OTHER-COMMIT. What is it that is happening in this case? What kind of a speech act is being performed here? Is it really a speech act that we want to consider and account for? It is not a very common construction and would appear highly rude if expressed in a real conversation. It is most likely to occur in the situation where one of the participants has a great deal of authority over another, enough so that they feel able to express a commitment on their behalf. The best examples of this behaviour come from those in a close relationship such as in a marriage, or between siblings, or between parents and children, or in a highly structured institutional setting, such as in the army between a superior officer and a soldier. It is within these pairings that I think OTHER-COMMIT acts can be found, e.g.

---

[9] I will ignore interrogatives for the moment as I intend to treat them within the same framework as other prescriptions; I will show how they fit in later.

'George will go to the shops for you, I'm sure.'

I am not entirely sure whether these acts should be characterised in the same way as DIRECT and SELF-COMMIT – perhaps they should be treated as an indirect version of the DIRECT act? Although, this is not a very satisfactory solution as OTHER-COMMIT can be issued without the presence of the committee (the intended agent of the prescription). However this may turn out to be, I have included them in the discussion here for the sake of completeness, and to show how they might eventually be incorporated in my model. I have found no other literature that accounts for this strange phenomenon.

To return to the main discussion, it is after the performance of the initial focus-setter prescription in the first turn (one of the three options itemised above) that the options begin to propagate uncontrollably. If we vary the parameters in all their permutations for just three participants, there are a possible 54 different alternatives for the response:

2 different propositions (X and ¬X)[10]
x 3 different speakers
x 3 different addressees
x 3 different agents
= 54 different options

This would mean that after only three turns the number of possibilities becomes massive: 8,748 to be precise (3 initial acts x 54 x 54). It would be absolutely impossible to characterise all of these acts individually. Not only is this beyond the ability of any theory to specify, but also the problem only escalates with the addition of other participants, or of a greater depth of turn analysis. The equation that expresses the complexity of this exhaustive list of options is,

$$c = 3((2p^3)^{(t-1)})$$

where: **c** = the total number of combinations, **p** = the number of participants in the conversation and **t** = the number of turns. So, the total number of combinations, **c**, is given by the three initial acts, multiplied by the two propositional stances (X and ¬X) times the number of participants, **p**, cubed (to allow for all the different permutations of roles), multiplied by the number of turns, **t**, minus the initial one (because there are only three options for the initial act and we have already included these in the calculation).

---

[10] Note that there is a third option: that of expressing zero commitment to the proposition by saying 'I don't know…'. I shall disregard this possibility for now, but will discuss it further in Section 10.1.3.

Clearly each one of these combinations could not possibly have a unique speech act interpretation. So, the problem then becomes one of deciding what are the essential features and parameters that are needed to identify prescriptive speech acts uniquely. If in reality the possibilities are so endless, what happens, if we restrict the number of possibilities that we look at, if one of the uncharacterised combinations occurs? I justify restricting the kinds of prescriptions that can follow each other by drawing an analogy with sentence construction using a particular grammar. Given a list of words, one can generate any number of sentences of any length. But, if the sentence is ungrammatical, then it will almost certainly make no sense whatsoever[11]. Similarly, sentences can be endlessly long, but in general we reach a natural end given by acceptability and memory constraints. Although the grammar might legitimise endlessly long sentences, in practice these are increasingly unacceptable. These two observations about sentence production and interpretation equate to **competence** and **performance** constraints.

In just such a way, many of the combinations generated by the exhaustive listing of the different kinds of prescription that could possibly follow each other will make no conversational sense whatever. There is also a natural saturation point for discourse exchanges, when participants feel that a topic has been 'done to death'. There is, I am claiming, a grammar of discourse. There is an expected structure to the performance of speech acts and a limit to the length of conversational segments. If that structure is violated, then either the conversation has gone wrong in some way, or there is some other explanation for the behaviour.

So the question is, how can we restrict the characterisation of the prescriptive speech acts performed in conversation in such a way that they are functionally unique yet not so numerous as to be unmanageable (analogous to the equally useless case when words are categorised so uniquely that there is a one to one correspondence between the category and the word itself).

Let us therefore look at each parameter in turn and consider how we can limit its consequences on speech act performance and recognition. Firstly, what can be done to limit the number of 'turns' that we consider? If we think about it, after the initial speech act, each participant can only really agree to the prescription or not [12]. The possible play of options can be mapped out relatively simply in Figure 9.1.

---

[11] Note that it will make no sense in a different way than the production of a sentence which observes some kind of internal structure, but is incoherent because the senses of the words themselves are incompatible, e.g. 'The purple sky flaps wickedly under a sharp sea'. (This idea is closely related to the ideas put forward by Chomsky (1965) concerning language creativity.)

[12] Or, as mentioned previously, express no commitment either way.

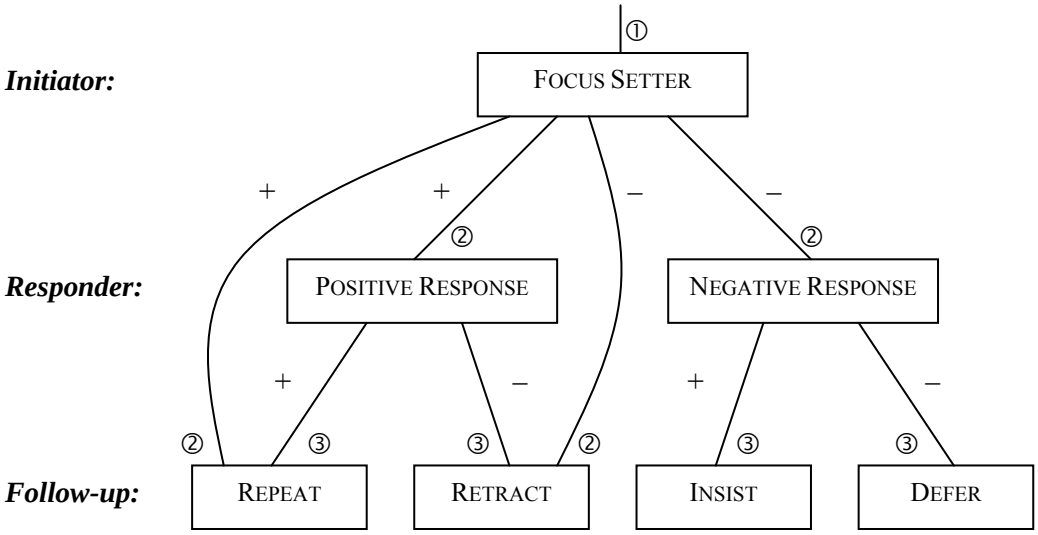


Figure 9.1 Seven basic speech act types.

The diagrammatic representation of speech act structure shown in Figure 9.1 is theoretically pleasing for a number of reasons. It fits it very well with ideas proposed by conversation and discourse analysts (e.g. Tsui 1994: 52) concerning the three-move sequence in which speech acts are performed within an exchange: initiating, responding and follow-up acts (as discussed extensively in Chapter 5). The structure also fits both types of speech act: descriptive and prescriptive. Although they are realised in different ways for different types of act, they are fundamentally functionally related. This is intuitively very appealing as it opens the way for potentially unifying the two types of speech act under one over-arching theory. I have separated the analysis of these different speech acts because they deal with different speech entities; descriptive acts deal with participants' stances towards information while prescriptive acts deal with participants' stances towards proposed actions. This division is largely artificial. In real conversation there is an interaction between the two types that I have so far failed to capture in my model, but hope to account for later. For example, prescribed actions are often complied with in the physical world (by the agent carrying out the prescription) and their completion reported by the agent[13]. Clearly in such a case, stating that some action is completed is performing more than an assertion in the context. Similarly, bald statements of fact like, 'The door is open', are rarely intended to be passively accepted or rejected by a hearer, especially if the fact is already plainly obvious to him. Under these circumstances the hearer will ask himself the question 'What does the speaker want me to do about this information?'. This requires a level of inference that again I am currently not able to cover in my model.

---

[13] At a later stage, my speech act model will also have to be incorporated into a more general theory of action in order to be able to deal with the fulfilment of speech acts by acts other than those performed in speech, e.g. a command to close the door being acceded to by the physical act of closing the door.

Further proof that descriptive and prescriptive speech acts are closely related and have similar types of function in their underlying structure can be found in the use of some of the same illocutionary verb in both types. Compare and contrast the following:

Descriptive	**Prescriptive**
Agree that…	Agree to do…
Promise that…	Promise to do…
Concede that…	Concede to demands…
Insist that…	Insist on doing…
Accept that…	Accept to do…
Defer to opinion/knowledge	Defer to wishes (etc.)

Here illocutionary verb usage is functionally overloaded, and I suggest that this is possible precisely because they play similar kinds of roles at the point in the interaction where they occur, even though they are managing different kinds of content – see the tree diagram in Figure 9.1.

We can see from Figure 9.1 that our different stances are exhausted three 'turns' deep. The problem with this terminology is that I now mean 'turns' in quite a different way to the accepted definition in linguistics. That is why from now on I shall speak of three levels of **stances** instead of using the word 'turn', which has such a specialised meaning. Stances could be composed of almost any number of turns (in their proper linguistic sense), but generally they are minimised by reaching a point of equilibrium if possible.

The participants adopt a stance by the performance of a speech act, either an initiating or a responding one. It is the recognition of a person's state of stance that allows us to interpret subsequent speech acts and so also either reinforce or change a current stance. Stances correspond to the 'state of activity' of the co-ordinated activity (as discussed in Section 8.2).

The goal of most conversation is to get to the point where all participants' stances coincide – i.e. that everyone has either accepted or rejected the focus-setter, or come to some mutually acceptable compromise. It is for this reason that conflicting responses will usually require the generation of some kind of explanation for the conflict. In other words, because people in general tend towards co-operativeness, it is not good enough to say 'no', one has to say why one is saying 'no'. I shall write more about explanations in Section 10.1.6.2 when discussing extensions for the model, but mention it here in passing. The current context space will not be deemed 'closed' while different participants retain conflicting stances, unless participants have agreed to disagree and start a new (or related) topic of discussion, a new context space.

Although Figure 9.1 is an over-simplified representation of the kinds of interaction between different types of speech act (for example, any act can be repeated or retracted, including the type INSIST and DEFER), it does show the outline structure that underlies all interactions.

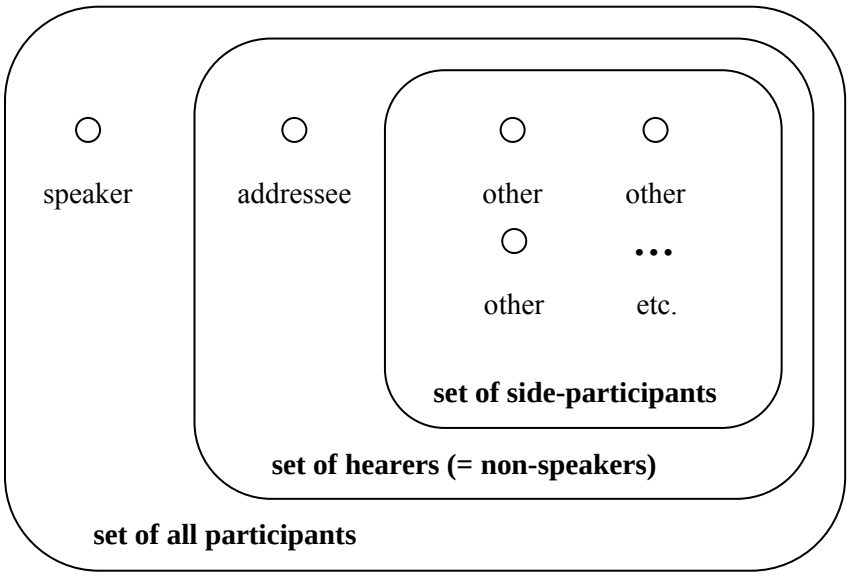


Figure 9.2 Ratified participants in a conversation in the model.

Having looked at restricting the 'length' or 'depth' of dialogue that the model will analyse for a particular speech act interpretation (by only considering the current stances of the various participants towards the focus-setter), I shall now aim to limit the number of participants that have to be considered for each speech act within a conversation. Ideally we want a model that will be flexible enough to encompass any number of participants within a conversation, without it becoming so complex as to be unworkable. We can do this by circumscribing the roles of the participants in every utterance. I shall argue that in any conversation there are actually only three types of participant that one can be: speaker, addressee or what is sometimes called the side-participant, but I shall call 'other'. In claiming this, I am not suggesting a different participant structure to that proposed in Chapter 2, Figure 2.2; just a slight modification. In a conversation between 'ratified participants' there is the basic divisions of role shown in Figure 9.2 for every turn.

If we take this structure and only consider the effect of these three types of roles in the interpretation of a speech act, then we can expand the model to include any number of participants without loss of expressiveness and without any increase in complexity. This is also

instinctively attractive as it mimics the participant structure to be found in grammar[14]. So, what we have per utterance is three different alternatives:

1st Person	**2nd Person**	**3rd Person**
Speaker (petitioner) – the 'owner' of the current utterance. Usually one person, very rarely more.	Addressee (respondent) – the expected next speaker. At least one person, except in the case of a monologue. Could be more than one if it is unclear who the addressee is intended to be, in which case any hearer might take on this role.	Other (witness) – the 'eavesdropper' of the utterance. Not expected to contribute to the current subject, but could do. This could be any number from none (in the case of a duologue) upwards.

All three parameters are subject to change after every utterance event. So the question is, how can we reconstruct what these parameters were for a past utterance from what we store and yet keep the possible number of permutations to a minimum? I will suggest a temporary solution to this problem in the following discussion.

Having restricted the roles of participants to three, we now turn our attention to the combination of these three parameters for each prescription. There are two further ways in which we can simplify the representation and thereby restrict the number of possible 'legal' speech act combinations.

The first is if we say that the agent of the prescription is always fixed from the focus-setter. This will immediately cut down the potential context spaces and will make sense of the feeling that who the agent is, is intrinsic to the prescription that we are currently discussing in the context space. That is not to say that speakers will not change the agent of the prescription during the course of the conversation – they often do, especially if their original proposal is rejected (see Section 9.2.6 for an example). But I suggest that what happens in this case is that a new context space is opened and the resolution of the prior context space will depend upon the resolution of the new one. For example, if your addressee refuses to accede to your direction, you may take it upon yourself to perform it instead, or undertake to ask someone else, or indeed ask one of the other participants directly. This will in effect close the old context space for the prescriptor (because one cannot be committed to conflicting propositions). Some sequences of replacing one agent for another will be inadmissible in a conversation because they would not make any sense – for example, if you have already expressed a directive act for **A** to do X, then

---

[14] At least for the languages that I am familiar with: English, Spanish, French, Danish and Latin.

you are unlikely to follow this by saying that **B** should not do X[15] (because the latter is entailed in the former), whereas if you then said that **B** should do X, you are either being inconsistent, or you are in some way changing your mind, unless X is the kind of action that requires two people. Here we could start to lose some of the simplicity of the representation that has been described so far because we might have to define co-actors, or multi-agents. For now I shall leave this issue and be content to note that it is a thorny one that will have to be addressed in the future. For now I shall treat the replacement of one agent for another as discretionary and as a contracted version of removing a prior commitment to one prescription and adding a commitment to a new one (hence the opening of a new context space).

I am not entirely happy with this solution, as it seems to constrain the options overly. Can we really see this as a wholly new context space? Or is it a nested context space – after all, the focus is still the same proposition, it is only the agent that has changed. I have decided to make it a new context space because it affects the recognition of any consequent speech acts. This will have to be reviewed in future work.

The second way of restricting the number of combinations of acts is by not indicating explicitly who is the addressee of the prescription. Again, this is not ideal as there is no doubt that in conversations of more than two participants the addressee of a prescription is often specified and known to all the participants. If the non-addressee responds, then this will often generate a response of something like 'I didn't ask you', or 'Who asked you?' or 'I wasn't offering it to you', etc. I justify omitting who is the addressee explicitly because it is mostly possible to infer who is the intended addressee in the case of directives from the information stored in the context spaces by the intended agent. This will not work with commissive acts because the agent is oneself, but often this kind of act is performed without specifying the addressee[16]. Questions are also quite commonly performed without a specified addressee; perhaps this is to maximise the chances of obtaining a satisfactory answer. In such cases all the hearers will be addressees, and the prescription can be accepted or rejected by more than one person. If there is a specified addressee and an unspecified 'other' accepts the commission, then that would be odd (much

---

[15] You might in fact perform these apparently equivalent acts for emphasis, or to spell out the consequences in detail, but it would be unusual. E.g. 'A do X – that means that B you won't have to do X'. Note that there is a big difference between prescribing that someone should do something and prescribing that someone should not. E.g. 'A don't do X – B will do it'. In this case, the second part of the utterance is not entailed in the first, so in fact we are adding a new prescription to the context. The form of this new prescription is that of an OTHER-COMMIT act, because the speaker is still addressing himself to A, not B.

[16] Perhaps because it will be clear in these cases that the addressee will be the person to most benefit from the completed action. If there is no definitive beneficiary to the commissive, then an addressee will not be specified anyway.

more so than if the non-addressee of an act of direction supports or countermands someone else). If we do away with the addressee, we would also cease to be able to account for the case of OTHER-COMMIT[17] (as discussed above), because we would no longer be able to tell the difference between this act and a direction. This is not utterly unacceptable, because in many ways one feels that this type of act is an indirect direction anyway. In the future, the addition of an explicit parameter for 'addressee' might help to distinguish more acts, but for the present, for simplicity's sake, we will assume that it is unnecessary.

By using these constraints, we end up with a much lower level of complexity; after one of the two different initiating acts (DIRECT or SELF-COMMIT, since we are now unable to distinguish OTHER-COMMIT), for three participants, there are only six different alternatives to follow:

	2 different propositions (X and ¬X)
x	3 different speakers
=	6 different options

This would mean that after three 'turns', or stances, the number of possibilities is only 72 (2 initial acts x 6 x 6). The equation that expresses this reduced complexity is,

$$c = 2((2p)^{(t-1)})$$

where: **c** = the total number of combinations, **p** = the number of participants in the conversation (now a constant of 3) and **t** = the number of 'turns', or stances (also a constant of 3). So, the total number of combinations, **c**, is given by the two initial acts, multiplied by the two propositional stances (X and ¬X) times the number of participants, **p** (which denotes the number of possible following speakers), multiplied by the number of turns, **t**, minus the initial one (because there are two options for the initial act and we have already included these in the calculation).

Although this may still seem like a high number of distinct speech acts to characterise individually, we have certainly improved on the original problem, and besides, I will show that this number can be cut down still further because many of the combinations are repetitious. As I hope will become clear, the number of viable combinations is reduced after every move.

---

[17] Although for non-initiating acts, we can tell OTHER-COMMITs from what is stored in the context.

So, if we can indeed specify all the possible permutations of the context space then I am claiming that speech acts are uniquely identifiable by the conditions that must obtain in the conversational context in order for their successful performance. To illustrate this, first let us look at the simplest type: the model of descriptive speech acts.

9.2.2 Descriptive Speech Acts

Descriptive speech acts are simple to define in that they consist of expressions of the various participants' attitudes to the truth or falsity of a proposition. As I mentioned in the previous section, they can be enumerated fairly straightforwardly. There is no need to specify an agent (because there is no prescribed action), and at the moment I do not distinguish between the addressee of the assertion and an 'other' hearer. This is because the addressee is very often left unspecified in *assertive* exchanges, even when there is more than one choice of following speaker. This does not have an effect on the identification of the act. In Table 9.1, I list the preconditions that must hold in the most recent context space in order for the current proposition **X** to 'count' as a specific descriptive act. (Use the key given below to interpret the table in this section and in those following.)

KEY to Context Space Tables		
⏵	=	future indicator, read as 'will bring it about that'
⏴	=	past indicator, read as 'has brought it about that'
∅	=	null or empty context space
X	=	the propositional content of the current utterance
SA	=	the speech act referred to by the speech act interpretation
¬	=	'not', negation[18]
s[*peaker*]	=	the current speaker
a[*ddressee*]	=	the current addressee
o[*ther*]	=	any one member of the set of participants, such that $o \neq s$ and $o \neq a$
h[*earer*]	=	any one member of the set of participants, such that $h \neq s$
p[*articipant*]	=	any one member of the set of participants
_	=	any other commitment to the proposition (including none, ∅)
❶, ❷	=	indicates chronological precedence, where ❶ is more recent than ❷.

[18] I have assumed that ¬X means: 'the negation of the current proposition X'. This could itself be negative, in which case ¬X could be positive (in precisely the case when X stands for the term (¬X), so that its negation ¬(¬X) would be simply X). In fact, X is just as likely to be positive as negative.

Context Space *speaker*	*hearer*	***s* SAYS X**
∅	∅	ASSERT
❶ X	_	REPEAT (SA)
❶ ¬X	_	RETRACT (SA)
∅	X	AGREE (SA)
∅	¬X	DISAGREE (SA)
❷ X	❶ X	ACKNOWLEDGE (SA)
❷ X	❶ ¬X	INSIST (SA)
❷ ¬X	❶ X	CHANGE-MIND (SA)
❷ ¬X	❶ ¬X	CONCEDE (SA)

Table 9.1 Assertive speech acts.

Later, in Sections 9.2.4 to 9.2.6, I will show that this is not the end of the story as far as the characterisation of descriptive speech acts are concerned. The list in Table 9.1 will have to be amended if we are to account for other kinds of speech act, such as recognising speech acts as answers to questions for example, which are sometimes, in the right context, descriptive responses to requestive acts. I shall come back to this point later, but for now we content ourselves with a clearer, if incomplete representation.

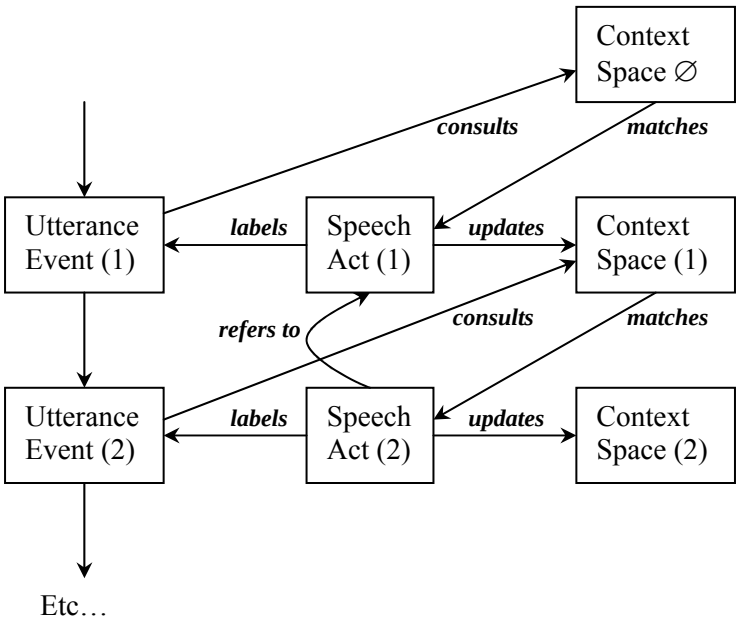


Figure 9.3 The working of the model.

Before discussing some of the issues that are raised by this representation, it might be helpful to clarify how to interpret Table 9.1 with an example. After every utterance in the current topic of focus, the model identifies the speech act performed by looking for a match with the

preconditions, what I call the **context space** (or in the terminology of Chapter 8, the state of activity), at the time of the utterance's performance. Once the speech act is identified, the context space will be updated; these are the **effects** of the speech act upon the context space (Figure 9.3 shows this diagrammatically). This idea is related to other work carried out on speech acts in computational linguistics (see Chapter 6 for details). These updates are crucial, as they will affect the subsequent performance of the model.

Note that I do not explicitly list the effects of an utterance on the context in Table 9.1. The reason for this is that the updates are the same each time. The context space is always updated by the replacement of the speaker's prior commitment by the current one, including the speech act interpretation. Only the 'owners', the speakers of the utterance, are entitled to retract what they say from the context. When this happens, a commitment to the proposition's negation is also added to the speaker's context space; the one subsumes the other, because you cannot be inconsistent with yourself; you cannot be committed to both X and ¬X at the same time (unless perhaps you suffer from a split personality). With the subsumption of the negated proposition, it is necessary to check to see whether any other speech acts can now be identified (see Figure 9.4).

(1)	**Mandy**:	oh, it's raining again.	ASSERT
(2)	**Julie**:	no it isn't.	DISAGREE (1)
(3)		oh, yes it is	RETRACT (2) & AGREE (1)

This conversation is represented diagrammatically in Figure 9.4:

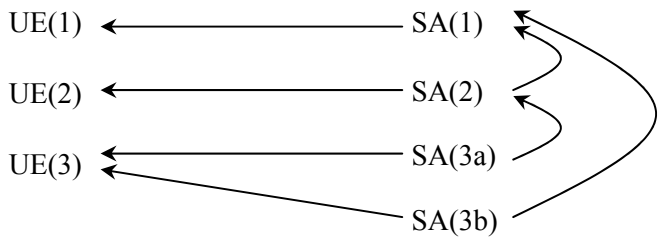


Figure 9.4 An example of context space consistency, updates and multiple speech acts.

In Figure 9.4 we see an example of how a single utterance might perform two acts at the same time. This is even clearer when there are three speakers, when a person could be agreeing with one participant and disagreeing with another at the same time. This poses a real problem for those (Searle 1979, Tsui 1994 for example) who insist that an utterance performs one and only one speech act in the context of a conversation.

Context spaces remain open in general until, ideally, participants achieve (perceived) consensus. Inconsistencies between different speaker's context spaces highlight conflict. This may in the future have important repercussions on the generation of appropriate speech acts and on whether a subject is pursued or not; the resolution and homogenisation of context spaces is a general goal of conversation.

Note that the context space is **not** the full conversational history, but the participant's **current** attitude and commitment to the proposition under focus. The full focus history is the sequence of context spaces, or the sequence of 'focuses', that lead to the current context space. This idea is very like the analogy of chess that I included in Chapter 8. It is the speech act that maps one context space (state of activity) onto another. One can recreate the conversation either from the sequence of speech acts, or from the sequence of context spaces (see Figure 9.5).

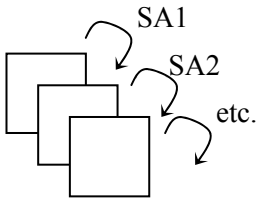


Figure 9.5 The sequence of context spaces.

The speech acts that come after the focus-setter (in the case of descriptive speech acts, the focus-setter is ASSERT) have the property of being, in effect, compound speech acts which depend on the performance and interpretation of a prior speech act (SA) for their identification. They are referring speech acts: they refer back to a previous utterance event in the conversation[19]. As such, rather than explicitly stating the content of the utterance event in the parameter list of the speech act as in the case of the focus-setter, a label indicating the utterance event (and speech act) to which the speech act interpretation of the current act refers is included. This has manifold representational advantages:

(1) It creates a linked chain of references that can be retraced at need;

(2) It avoids the need to list all the options for the speech acts REPEAT and RETRACT, because they can refer to different speech act interpretations simply by referring to the content of the utterance event.

(3) Following on from (2), it allows us to cut down on the number of distinct speech act combinations (as calculated in Section 9.2.1) by allowing the same act to refer to more than one other type of speech act. With a smaller set of speech act types one can build up a large

---

[19] Analogous to backwards-communicative function in the DAMSL mark up scheme.

variety of combinations, e.g. the speech act REPEAT can refer to ASSERT, but depending upon the state of the context, it can also refer to any other speech act type, descriptive or prescriptive.

(4) It allows the same utterance event to have different speech act interpretations depending upon which of the previous utterance events it refers to.

(5) Speech acts are clearly also referential in conversation, which explains the use of elliptical answers such as 'Yes' (to the assertion 'It's a beautiful day'), or 'Five o'clock' (to the question 'What's the time?'), for instance.

(6) It avoids the complication of such speech acts like DISAGREE, which have a negative built into the label. Passing the content in the parameter, DISAGREE (X), would suggest that the speaker is saying ¬X, whereas, non-intuitively it really means that the utterance content X is classed as a disagreement in the current context. By using DISAGREE (SA), it makes it easier for a reader to think about the representation in terms of: 'the current utterance is interpreted as an act of disagreement with the content of the utterance event, SA'. It is a short form for the nested representation: DISAGREE (*s*, ASSERT(*h*, X)).

Although the act of repetition will under the current development of the model have no visible effects on the state of the context, in fact this will not be the case. Repeating a previous speech act will alter the current focus of attention in the conversation, and might also have some effect on the strength of commitment by the speaker to the content of the utterance. If this model were to be extended to deal with these phenomena then the effects of the act REPEAT would be of more significance.

There is a problem with the representation here, because if I say that X, then follow it up immediately with ¬X, one would assume that this is a retraction of the original proposition. This is not only an act of retraction, however, it also commits the speaker to the truth of ¬X. This new commitment would be added to the list of the speaker's commitments. In order to avoid having contradictory commitments, the one would have to replace or subsume the other. So an act of retraction does not just remove the commitment from the context space, but inevitably also adds its negation to it also.

If a retraction of some description occurs later in the conversation (with perhaps some intervening sequences of speech acts in between), the interpretation might be quite different. A retraction almost always requires some kind of an explanation on the part of the speaker to justify it, but especially so under these circumstances. Without an appropriate explanation backing up the retraction, the speaker might simply be deemed inconsistent. This is similar to the case when a speaker later says Y, which implies ¬X. It is inconsistent with X, so either the speaker did not really mean Y now, or did not really mean X in the first place. Although the

model in its simplest state does not handle cases that involve inference, in Section 10.1.6.2 I will explain how they might be incorporated later.

There are also issues of permanency versus non-permanency to consider when looking at speaker consistency. Some things have states of being (like the weather for instance), so if one says X ('It is raining') and later follows this up with ¬X, this does not necessarily imply contradiction unless X is anchored in time (e.g. 'It was raining at 6.00pm'). This kind of temporal logic might be integrated into a more complex model at a later date.

For the purposes of identifying legal sequences of speech acts, I have assumed that speech acts made in the negative are processed in exactly the same way as those made in the positive (i.e. if I ASSERT that X, X can stand equally for a well-formed proposition in the negative such as 'It is not raining' as it can for one in the positive such as 'It is raining').

At first I had thought that negative speech acts are only ever used in response to some previous positive speech act, and that the performance of an initiating speech act in the negative would be rather odd and would necessarily presuppose a context where the responder was already committed to some claim or action (and that therefore the context was not truly empty). For example, imagine starting a conversation with:

'Don't go shopping tonight.'

This would immediately imply that the speaker had good reason to suppose that the addressee intended to go shopping tonight. If the addressee had already stated their intention to do so, then this would be reflected in the context, so perhaps someone else had informed the speaker of the addressee's intended action (which would explain a response such as "Who told you I was going shopping?" or "How do you know I'm going shopping?"). Alternatively, the addressee habitually goes shopping on this night of the week and the speaker is aware of that fact (in which case, perhaps again the context is not entirely empty with regards to this proposition). Howsoever, the implication of the performance of this speech act of forbidding (or 'commanding that not') is that all things being equal, in the normal course of events the addressee would have carried out the action of going shopping tonight.

However, the performance of <u>any</u> speech act, positive or negative, carries with it the assumption that in its very performance the context would be changed. If a conversation were to be started with (the slightly rude) converse of the above example:

'Go shopping tonight.'

The speaker would just as equally have to assume that the addressee was not intending to go shopping tonight of their own volition. There are some speech acts in fact that are more likely

to be performed negatively than positively. The act of warning for example is often performed in this way.

> ‘Don’t touch that!’ (because it has just been painted)
> ‘Don’t disturb her!’ (because she’s in a bad mood)
> ‘Don’t use the phone!’ (because I’m on the Internet)

Though it is not always the case that a negative directive can be interpreted as a warning – this is also determined by some semantic and logical processing, and by whether compliance with the content of the direction is in the interests of the speaker or of the addressee. I shall be returning to this point in later discussion.

So, if we accept that negative speech acts can be used in just exactly comparable circumstances as positive speech acts, this would be theoretically very elegant as they could be treated in the same way without a need to specify different rules for negative speech acts.

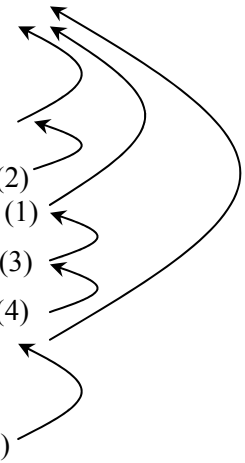
In the model depicted in Table 9.1, I have distinguished the act of acknowledgement from that of repetition because of a feeling that there is a difference between mere repetition and a response to a hearer’s most recent utterance. I am not entirely convinced that this is necessary, as it makes no functional difference to the working of the model. Agree and acknowledge are very similar as well. Insistence is also closely related to repetition, but in this case, it is clear that in order to insist that something is true, there must be some hearer who has claimed that it is not. I believe that this is one of the strengths of the three-level state-trace model described in this dissertation, as it is the only attempt to define speech acts that rely on the performance of more than one previous speech act for their interpretation.

Similarly, if we look at the ‘negative’ speech acts, we could almost certainly conflate the act of changing one’s mind and the act of retraction. As I have already explained, when you retract a commitment to the proposition, you invariably become committed to its negation. It is not clear to me that these are differentiated in real conversation, so it should be possible to get rid of this distinction and simplify the model further. However, I include all of the options here for the sake of consistency and completeness.

To bring the discussion on descriptive speech acts to a conclusion, I include an example below with its corresponding speech act analysis. Notice that in this example, the speech act labels seem to oversimplify what is actually happening in the conversation. For example, the utterances (2) and (3) are arguably one act of partial agreement, which is indicated by the cue phrase ‘yeah, but…’. I shall be looking at this phenomenon in more detail in Section 10.1.5. For now I would just like to note that the speech act model in its current form provides a framework for the development of a more complex model later on.

Example of the Flow of Conversation from a ASSERT Focus-Setter:

(1)	**Nigel**:	well the whole thing about fly-on-the-wall TV documentaries is that it takes time for people to adjust to it.	ASSERT
(2)	**Mandy**:	(pause) yeah,	AGREE (1)
(3)		but I've only got three hours (laughing) of tape.	RETRACT (2) DISAGREE (1)
(4)	**Nigel**:	yeah	CONCEDE (3)
(5)		no that that'll pro- probably be enough I'd have thought	RETRACT (4) INSIST (1)
…			
(6)	**Nigel**:	nah, I mean see already we've just been talking naturally there.	REPEAT (5)



In this section, I have presented a preliminary model for descriptive speech acts. I shall now provide an account of the more difficult cases: prescriptive speech acts.

9.2.3 Prescriptive Speech Acts

Prescriptive speech acts (which cover directive and commissive utterances, as well as being strongly related to interrogative or, as I shall call them, requestive utterances) are much harder to define. They are more complex because in order to identify them exhaustively, one must take into account an extra parameter: the *agent* of the prescription (for an action) contained in the content, as we have already discussed at length when describing the development of the model in Section 9.2.1.

As discussed earlier in this chapter, the difference between a directive and a commissive speech act is quite simply that in the case of a directive, the addressee is the intended agent of the action prescribed in the utterance, whereas in the case of a commissive, it is the speaker. I treat these two types of prescriptive act under the same heading not only because of their similarity of function, but also more importantly because they work in conjunction with each other. Directive and commissive speech acts tend to be interlaced in a conversation for obvious reasons, with a directive type following a commissive and vice versa.

Prescriptive types of speech acts share the property that the action prescribed in the utterance should be carried out at some future point in time. So in the case of a directive utterance, if ***s*** **SAYS** ***a*****: ▸ X**, this should be interpreted as ***s*** **DIRECTS** ***that the addressee should bring it about that the propositional content X holds true in the context***, e.g. 'Shut the door' is represented by ***s*** **SAYS** ***a*****: ▸ The door is shut**. Similarly, in the commissive case, if ***s*** **SAYS** ***s*****: ▸ X**, this should be interpreted as ***s*** **SELF-COMMITS** ***that the speaker should bring it about that the propositional content X holds true in the context***, e.g. 'I'll shut the door' is represented by ***s***

SAYS *s*: ▸ The door is shut. In Tables 9.2 and 9.3 I list the preconditions that must hold in the most recent context space in order for the current proposition ***p*: ▸ X** to 'count' as a specific prescriptive act. Use the key given in Section 9.2.2 to interpret these tables.

Context Space *speaker*	*addressee*	*other*	***s* SAYS *a*: ▸ X**
∅	∅	∅	DIRECT
❶ *a*: ▸ X	_	_	REPEAT (SA)
❶ *a*: ¬ ▸ X	_	_	RETRACT (SA)
∅	*a*: ▸ X	∅	ACCEPT (SA)
∅	*a*: ¬ ▸ X	∅	DECLINE (SA)
∅	∅	*a*: ▸ X	SUPPORT (SA)
∅	∅	*a*: ¬ ▸ X	COUNTERMAND (SA)
❷ *a*: ▸ X	❶ *a*: ▸ X	∅	ACKNOWLEDGE (SA)
❷ *a*: ¬ ▸ X	❶ *a*: ▸ X	∅	DEFER (SA)
❷ *a*: ▸ X	❶ *a*: ¬ ▸ X	∅	INSIST (SA)
❷ *a*: ¬ ▸ X	❶ *a*: ¬ ▸ X	∅	CHANGE-MIND (SA)

Table 9.2 Directive speech acts.

In Table 9.2 there are two speech acts – SUPPORT and COUNTERMAND – that can only be identified by considering the context space of *other* participants. This was a late addition to the model, which makes it more powerful, but means that the effects on the recognition of subsequent acts have yet to be fully explored.

Example of the Flow of Conversation from a DIRECT Focus-Setter:

(1)	**Nigel**:	's ah yep, going now... yep.	ASSERT
(2)	**Mandy**:	now forget about it.	DIRECT
		(laughs)	
(3)	**Nigel**:	okay,	ACCEDE (2)
(4)		right,	REPEAT (3)
(5)		it's forgotten.	REPORT (4)[20]

The example of a conversation from a DIRECT focus-setter given above demonstrates some of the power of a representation built on structure and function. The model behaves in exactly the

---

[20] I shall ignore the REPORT speech act for the moment, but will account briefly for this extra speech act in Section 9.2.5. Note that the reference of this act links back to the REPEAT act; however, surely it should link to the focus-setter or response. Instinctively we would like to say that it is a report of an act of accession.

same way at this level of analysis irrespective of whether the directive is literal or non-literal, as it is in this case. It does not matter at the structural level that the directive is impossible to comply with (because the act of forgetting is not willed), the participants follow the speech act structure regardless. In fact, this could be the very reason why such a conversation is perceived as slightly humorous: because it flouts the expectation of a rejection of the directive.

In order to follow the example dialogue above, one also needs to consult the commissive speech acts as laid out in Table 9.3.

Context Space			***s* SAYS *s*: ▸ X**
speaker	*addressee*	*other*	
∅	∅	∅	SELF-COMMIT
❶ *s*: ▸ X	_	_	REPEAT (SA)
❶ *s*: ¬ ▸ X	_	_	RETRACT (SA)
∅	*s*: ▸ X	∅	ACCEDE (SA)
∅	*s*: ¬ ▸ X	∅	REFUSE (SA)
∅	∅	*s*: ▸ X	ACCEDE (SA)
∅	∅	*s*: ¬ ▸ X	REFUSE (SA)
❷ *s*: ▸ X	❶ *s*: ▸ X	∅	ACKNOWLEDGE (SA)
❷ *s*: ¬ ▸ X	❶ *s*: ▸ X	∅	DEFER (SA)
❷ *s*: ▸ X	❶ *s*: ¬ ▸ X	∅	INSIST (SA)
❷ *s*: ¬ ▸ X	❶ *s*: ¬ ▸ X	∅	CHANGE-MIND (SA)

Table 9.3 Commissive speech acts.

The addition of *other* participants in the model shown in Table 9.3 will not help to distinguish extra speech acts in the case of commissive speech acts (note the repetition of ACCEDE and REFUSE). This is because it is not important which participant it was who prescribed that the current speaker do X in the context.

Example of the Flow of Conversation from a SELF-COMMIT Focus-Setter:

(1)	**Nigel**:	I'm just gonna turn the volume up on it.	SELF-COMMIT
(2)	**Mandy**:	okay. (pause)	ACCEPT (1)
(3)		I don't think it makes any difference actually, to the input.	RETRACT (2) DECLINE (1)
(4)	**Nigel**:	right.	DEFER (3)

The speech acts defined in Tables 9.1 – 9.3 are compared in Table 9.4. Notice the similarity between the descriptive and prescriptive models. This is encouraging, as one would like to think that the structure of interaction has an underlying functionality in common. The use of in

some cases identical names reflects that in fact these labels are irrelevant and, to a large extent, arbitrary. I have chosen different labels (in the cases of *focus-setter* and *response* labels in particular, highlighted in bold in Table 9.4) for the sake of clarity, but it should be clear that functionally these speech acts perform in the same way, but on different conversational entities. I have used the same speech act labels for REPEAT, RETRACT, ACKNOWLEDGE, DEFER (labelled CONCEDE in the case of assertives), INSIST or CHANGE-MIND, because it makes sense to say that they can refer to any speech act. As I have mentioned previously, some of the speech acts are strongly related to, and might even be redefined in terms of, each other (e.g. INSIST could perhaps be deemed a special case of REPEAT; however, it is only by using a state-trace approach that these two acts can be distinguished).

Type (from Figure 9.1)	***Assertive***	***Directive***	***Commissive***
Focus-setter	**ASSERT**	**DIRECT**	**SELF-COMMIT**
Repeat	REPEAT	REPEAT	REPEAT
Retract	RETRACT	RETRACT	RETRACT
Positive Response	**AGREE**	**ACCEPT**	**ACCEDE**
Negative Response	**DISAGREE**	**DECLINE**	**REFUSE**
–	–	**SUPPORT**	**ACCEDE**
–	–	**COUNTERMAND**	**REFUSE**
Acknowledge	ACKNOWLEDGE	ACKNOWLEDGE	ACKNOWLEDGE
Defer	**CONCEDE**	DEFER	DEFER
Insist	INSIST	INSIST	INSIST
Change-Mind	CHANGE-MIND	CHANGE-MIND	CHANGE-MIND

Table 9.4 A comparison of descriptive and prescriptive speech act labels.

I have spent some time going through the development of the model for prescriptive speech acts earlier in the chapter in Section 9.2.1, and furthermore, much of the discussion concerning descriptive speech acts is relevant to prescriptive speech acts also; I will not repeat this discussion unnecessarily here. In Tables 9.2 and 9.3 in this section I have given the outline conditions for the performance and recognition of basic directive and commissive speech acts, as well as some simple examples. From these, it should now be clear how the model is intended to work. As this is only a preliminary representation, I do not claim that the specification of each act is entirely complete. My intention is to provide a means of understanding the idea as a whole.

Rather than dwell further on details, I will now turn my attention to a problematic, yet ubiquitous, set of speech acts that can perform many roles in a conversation at once. These are namely those performed largely in the interrogative mood: questions. In the next two sections I will present a discussion of the problem and will suggest that these kinds of speech acts, which I

shall term *requestives*, can be dealt with within the same framework as the other, perhaps more straightforward, speech act types: assertive, directive and commissive.

9.2.4 Requestive Speech Acts

I have separated out requestives – those speech acts performed in the interrogative mood – from prescriptives, as I will argue that they perform different, albeit related, kinds of acts. In the following sections, I shall be looking at the arguments and addressing whether questions should indeed be treated in this way, before showing how they fit into the state-trace model.

9.2.4.1 The Problem of Interrogatives

Are interrogatives a separate family of speech acts (indicating a deference to the addressee's opinion or wishes) altogether, or do they belong to the prescriptive type? It is not at all clear that they should be dealt with either separately or inclusively: the problem is illustrated in Figure 9.6 (the dotted line shows the alternative inclusive interpretation).

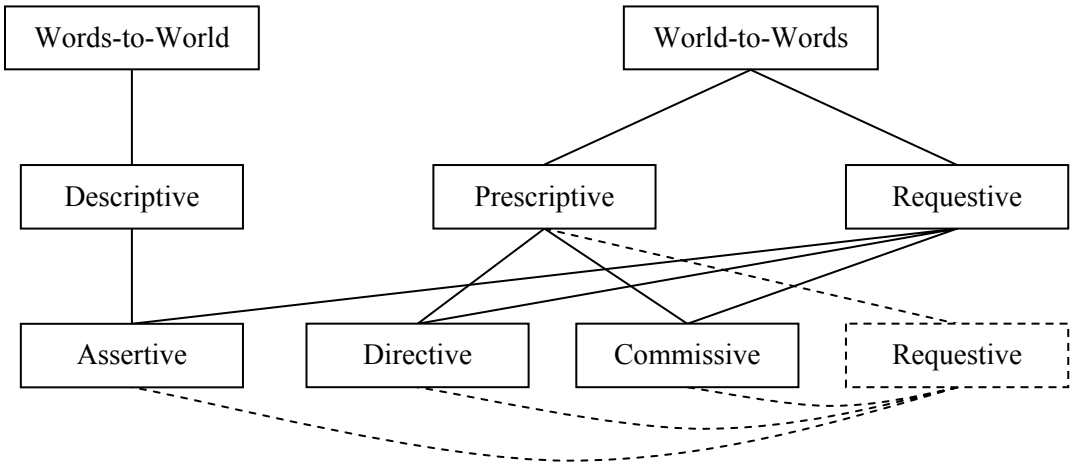


Figure 9.6 Breakdown of different types of speech act.

Lyons (1977) argues that questions are usually performed with the expectation of an answer, and that, as this expectation is conventional in nature, it is independent of the illocutionary force of the question. I would in fact go further and say that almost any utterance holds the expectation of an answer; the difference is that interrogatives explicitly prescribe an answer conventionally. In this sense questions are very much related to prescriptive speech acts, and should come under the world-to-words direction of fit. Again though, this is not immediately obvious. For example, how is the world to be made to fit the words of a question such as: 'Is the door open?'? Is it a short form of: the speaker prescribes that the addressee RESOLVE whether 'the door is open' is true or false'? If it is, then the requestive case works in a very similar (if not equivalent) way to prescriptives. The only difference is that whereas

prescriptives prescribe an **action** from the addressee, requestives prescribe a **speech action** to resolve either a descriptive or prescriptive content.

One way of dealing with questions is to say that the possible answers range over a set of propositions given in the utterance, and the question itself is a request for some kind of resolution. Treating questions in this way would mean that, depending on the kind of question asked, the underlying prescriptions are those given in Table 9.5.

	Type of Question	***Prescription***
REQUEST-ASSERTIVE	*WH-*	*s* prescribes *a* to: RESOLVE (P(*x*): $x \in \{set\}$)
	Alternative	*s* prescribes *a* to: RESOLVE ($\{P_1, P_2, \ldots, P_n\}$)
	Yes/No, Tag, Declarative	*s* prescribes *a* to: RESOLVE ($\{P, \neg P\}$)
REQUEST-DIRECTIVE	*WH-*	*s* prescribes *a* to: RESOLVE (s: ▸ P(*x*): $x \in \{set\}$)
	Alternative	*s* prescribes *a* to: RESOLVE ($\{s$: ▸ P_1, s: ▸ P_2, …, s: ▸ $P_n\}$)
	Yes/No, Tag, Declarative	*s* prescribes *a* to: RESOLVE ($\{s$: ▸ P, s: ¬ ▸ P$\}$)
REQUEST-COMMISSIVE	*WH-*	*s* prescribes *a* to: RESOLVE (a: ▸ P(*x*): $x \in \{set\}$)
	Alternative	*s* prescribes *a* to: RESOLVE ($\{a$: ▸ P_1, a: ▸ P_2, … a: ▸ $P_n\}$)
	Yes/No, Tag, Declarative	*s* prescribes *a* to: RESOLVE ($\{a$: ▸ P, a: ¬ ▸ P$\}$)

Table 9.5 Questions as requests for resolution.

This would mean that questions are basically represented as a request to RESOLVE a disjunction of their answers. The problem with this idea is that there are ways in which the addressee of a question may answer that cannot be predicted in the disjunction (not to mention other objections about this representation from the point of view of logic). Take for example the question 'Is it raining, sleeting, or snowing?'; the disjunction of answers to the question would be 'It is raining' or 'It is sleeting' or 'It is snowing'. However, the addressee might answer by saying 'What are you talking about? It's bright and sunny outside'. The speaker has answered the question by saying that it cannot be answered in its current form, and giving a justification for rejecting the original question. The reason for the failure of the question to signal the appropriate answer in its disjunction, is due to the failure of the presupposition that the disjunction contains as one of its elements the correct answer. How can this be represented?

There are other difficulties involved in representing questions. Some types of question are not as simply accounted for, or as easily fitted into the model. 'Why?' and 'How?' questions are considerably more complex than other *WH*-questions for example, as they require the ability to reason about propositions and provide explanations and motivations, rather than just the straightforward filling in of missing information. Even 'Who?' and 'Which?' questions will

typically require more complex answers, e.g.: ‘Who is that man?’, ‘That’s Ralph, Sheila’s husband’. We can see from the following examples that the representation of *WH*-questions is not quite as simple as we should like:

What is his name?
RESOLVE (His name is (x): $x \in$ {*set of names*})

When shall we meet again?
RESOLVE (We shall meet on (x): $x \in$ {*set of days*})

Who is the president?
RESOLVE (The president is (x): $x \in$ {*set of people*})

Where is Armley?
RESOLVE (Armley is in (x): $x \in$ {*set of locations*})

Which knife should I use?
RESOLVE (Use the knife (x): $x \in$ {*set of knife descriptions*})

Why are you frowning?
RESOLVE (I am frowning because (x): $x \in$ {*set of motives*})

How am I going home tonight?
RESOLVE (You are going home by (x): $x \in$ {*set of vehicles*})

I do not claim to give a full account of how these types of questions might be dealt with, and suspect that the representation adopted in Table 9.5, even for simpler types of questions, is strictly inadequate to explain both to what a speaker commits himself when asking a question (other than wanting to know the answer) and how the answer is to be formulated.

From the discussion so far, we might conclude that questions should indeed be treated as prescriptions to resolve uncertain information. Questions certainly have a prescriptive element about them in that they prescribe that the addressee should make the content of the question maximally informative. However, requestives **do** subtly differ from both descriptive and prescriptive speech acts. Requestives function as a means of performing incomplete other acts, which the speaker intends the addressee to complete. When using descriptive and prescriptive utterances the speaker has a definite, fully formed piece of information or a definite, fully formed proposal for action to convey to the addressee. Requestives on the other hand introduce a marker to show that the speaker requires that the addressee convey to him an specific piece of information or a proposal for action. In other words, questions are a way of reversing the direction of the flow of information; instead of from initiator to responder, from responder to initiator. Descriptive and prescriptive utterances presuppose that the initiator is implying ‘I tell you’, whereas requestive utterances presuppose that the initiator is implying ‘You tell me’. The balance of power is shifted so that the speaker, instead of providing a commitment, explicitly elicits one from the addressee. It is for this reason that I argue that requestives should be dealt with as a separate, albeit potentially equivalent, category.

9.2.4.2 Requests for Action vs. Directives

There is some debate as to whether requests for action and directives (usually labelled 'orders' or 'commands') should be characterised as different forms of the same category of act or separately as different categories of acts altogether. The majority consensus in recent linguistic theory (Katz 1977, Labov and Fanshel 1977, Bach and Harnish 1979, Searle 1979 and Leech 1983 for example), with which I shall be agreeing, is that these are in fact two strongly related means of performing the same kind of speech act. I have already looked at requests as indirect directives in Chapter 4. I will here briefly consider some of the arguments by Tsui (1994), who is one of the few linguists to argue against classifying these two types in the same way. She distinguishes between the following two utterances:

(1) 'Open the door'
a: ▸ The door is open *Directive*

(2) 'Would you open the door please?'
a: ▸ ACCEDE (*a*: ▸ The door is open) *Requestive*

Tsui calls (1) a directive and (2) a requestive (with the specific meaning of 'a request for action') and deals with them as different entities, from two entirely separate subclasses. I have indicated the difference between these two types by making the first a direct prescription that the addressee should bring it about that 'The door is open' is true in the (physical) context, and by making the second an embedded, indirect prescription that the addressee should bring it about that he ACCEDE to the prescription that he should bring it about that 'The door is open' is true in the (physical) context. This indirectness will have no effect on the way the response is interpreted. If the addressee accedes verbally to either of these utterances, he will be deemed to have committed himself to perform the action of opening the door; in the case of (2) above, the positive response will perform a dual role of answer and accession. In the instance that the addressee simply accedes to the prescription (embedded or not), there will be no perception that there is a step missing. So I will claim that the REQUEST-ACCEDE act stands in place of the DIRECT act and can be performed and responded to in precisely the same way; the two are functionally equivalent.

Tsui points out that using a command (in the imperative mood) as a method of issuing a directive is significantly different to a request in that it "does not indicate the speaker's doubt or query as to the addressee's carrying out the action, hence leaving the latter no option but to comply" (Tsui 1994: 93). In other words she argues that directives do not allow for the addressee to respond negatively, whereas requests for action allow for both a positive and a negative response. I disagree with this distinction. Except perhaps in certain institutional settings, a command can just as easily be refused as a request. Although conventionally the formulation in (2) is more polite than in (1), they are functionally performing in the same way.

It would be equally rude to respond to the underlying prescription with the curt 'No I won't' in both cases. In fact, if anything this refusal would be less expected or acceptable after the utterance of (2) precisely because of the polite formulation.

In intimate relations, commands are often issued instead of requests (see Conversation 1.1, Chapter 1) and refused just as easily. It would be odd if strangers used commands instead of requests, because strangers do not have either the familiarity or the authority to issue commands; strangers need to maximise their chances of being socially accepted by being as polite as possible. These points are important to the discussion because they illustrate that while the two types of act conventionally have appropriate and inappropriate usages within the social context, functionally they are almost entirely interchangeable.

Further evidence that Tsui cites to support the separate classification of directive and requestive utterances are the manner in which these two types are reported. For example:

'Would you please remove your glasses?' → He asked me to remove my glasses.
'Remove your glasses' → He told me to remove my glasses.

This evidence is not particularly strong, as it is quite conceivable that the speaker could report these acts in either way for either example. Far from backing up the notion that there is a functional distinction between directives and requests for action, I would suggest that this supports their similarity at the functional level.

9.2.4.3 Questions as Requests

Questions are problematic in many ways for speech act theory in general, and no less so for the model I have developed. There are a variety of opinions prevalent at this point in time concerning how questions should be viewed. In common with many others (Katz 1972, 1977, Gordon and Lakoff 1977, Labov and Fanshel 1977), I will take questions to be a special case of REQUEST. As I have already discussed in Section 9.2.4.1, requests are made of people to perform actions; questions are, I shall argue, requests to perform specific speech actions.

Type of Question	***Stenström***	***Tsui***
WH-	identification question	elicit: inform, commit, repeat, clarify
Alternative	?	elicit: inform
Yes/No	polarity question	elicit: inform, confirm
Tag	confirmation question	elicit: confirm, agree
Declarative	?	elicit: confirm, agree
Offers	permission request	offer / request for permission
Requests	action request	invitation / request for action

Table 9.6 Stenstöm's (1994) and Tsui's (1994) treatment of interrogatives.

In Table 9.6, I compare Stenström's (1994) and Tsui's (1994) treatment of questions. Note that both draw a distinction between questions and requests, whereas I will not. In fact I will assume that all questions are requests of some description or other (primarily for a response that includes a resolution of the missing information in the question), and that the distinction between questions and requests in the terminology of Stenström and Tsui is built into the difference between the descriptive and prescriptive contents of the question. I will also be arguing that the forms more commonly associated with mere information seeking questions (rather than requests for action) such as *WH-*, *Alternative*, *Yes/No*, *Tag* and *Declarative* questions are also used to elicit actions – see Table 9.7 for examples. In fact, the inclusion of *Offer* and *Request* under the heading of 'Types of Question' in Table 9.6 is, I believe, extremely misleading. In fact, *Offers* and *Requests* are generally realised by *Yes/No* questions.

Saddock (1974) gives two syntactic reasons why not all questions should be counted as requests. Firstly he claims that all requests can take the adverbial 'please'; however, there are questions that are used as indirect requests with which 'please' is not compatible, e.g. *'Don't you think you should please take out the garbage'. I am not convinced that this provides conclusive evidence for discounting questions as requests. If we rephrase this unacceptable sentence, we can produce an acceptable one: 'Please, don't you think you should take out the garbage?'. The unacceptability of the former example might furthermore be explained by the obligation implied by the use of the modal 'should'; contrast for instance *'Shouldn't you please take out the garbage?' and 'Couldn't you please take out the garbage?'.

If we accept that the adverb 'please' always indicates a request as claimed by Saddock, and we want to claim that all questions are a type of request, then there are other forms question that equally cause problems, e.g. *'Shall I take out the rubbish please?', which is generally interpreted as a request for permission (or an offer). This is because 'please' stands for 'if you please', which is a cue phrase indicating the recognition that the addressee is doing the speaker a favour. However, with commissive contents ('I shall take out the rubbish') this is not usually the case. So, I suggest that 'please' cannot always indicate a request, and certainly not in the

same sense as I intend to use the word 'request' (which is as a prompt for a speech act, sometimes also containing an embedded prescription for a physical act).

As an aside, the addition of 'please' can alter even straightforward *WH*-questions that prospect for information to having an underlying requestive force.

A: What time is it please?
B: [Yes.] It's half past twelve.

The use of 'Yes' at the beginning of the answer, while sounding slightly odd, is not that unusual in this situation. Is it just a means of gaining time while the responder consults his watch, or is it that the appended 'please' to the *WH*-question triggers a requestive interpretation (as if the responder had actually heard something like: 'Could you tell me the time please?').

The second syntactic objection that Saddock (1974: 90) has to treating all questions as requests is by reference to the maxim that all 'true' questions allow the pre-tag 'Tell me', whereas not all requests do, e.g. *'Tell me, take out the garbage, will you?'. Again this is not enough of a reason to reject the hypothesis that all questions are some sort of request at root. The question 'It's raining, is it?' would sound equally odd rephrased as *'Tell me, it's raining, is it?', which is the equivalent *Tag* question that asks for information rather than for action. It is, I believe, a peculiarity of the formulation of that particular *Tag* question that makes the addition of the pre-tag 'Tell me' ungrammatical, not the fact that it is also a request.

The problem could well be one of terminology: perhaps it is confusing to label questions 'requestives' because the word 'request' already has a specific meaning (i.e. request for action). In future this might be renamed to something like 'elicitatives' to help distinguish these, but for now I shall continue to use the term REQUEST to cover questions.

Difficulties also arise because the entities being elicited in what are traditionally labelled questions (asking for information) and requests (asking for action) are different. In the latter case, often the answer can perform more than one function at the same time, or can be performing functions at different levels. Lyons (1977) argues that it is for this reason that questions should be treated differently from requests. For example the answer 'No' to a *Yes/No* question provides an answer to the question content, whereas the answer 'No' to a *Yes/No* request is a refusal to perform the action requested. E.g.:

(3) 'Is the door open?'
(4) 'Open the door, please?' or 'Would you open the door please?'

In (3), the implicit request to provide information is acceded to in the very act of answering 'No'; in order to refuse the implicit request in (3) to provide information, the responder would

have to say: 'I'm not going to tell you'. In (4), the implicit request to provide a commitment to open the door or not is also acceded to in the act of answering 'No', while the embedded directive for action is refused in the same act.

Treating questions and requests in the state-trace model as the same kind of functional phenomenon, with the exception that they expect different types of speech act in reply (in the case of (3) a descriptive response, and in (4) a prescriptive response), provides an explanation both for why they perform differently and for how they can still be fitted elegantly into the same framework.

Tsui (1994) distinguishes questions that only elicit information in the form of a verbal response from requests that primarily elicit an action as a response, and only optionally a verbal response. She claims that:

> Questions have a different discourse function or consequence from requests and therefore they should not be subsumed under the latter. (Tsui 1994: 80)

In essence I agree with Tsui that what she calls requests are indeed different to what she calls questions and do have different discourse functions, but only because it is the type of element requested that is different. Tsui distinguishes information questions by calling them *elicitations*, "since the category 'question' is vague and ill-defined and cannot be subsumed under either requests or directives…" (ibid.). She seems to have discarded the simple interpretation that it is the force of questioning, the interrogative mood, that converts the directive into a requestive and the informative into an elicitation (Tsui's terminology).

Treating questions as requests is a singularly attractive method of dealing with them as it is based on an analysis of discourse function rather than syntactic form. It chimes in well with my analysis of other types of speech act in this respect. I have not hitherto discussed how little real correlation there is between the traditional moods declarative, imperative and interrogative and the speech act types descriptive, prescriptive and requestive. For example, commissive speech acts (which are prescriptive in nature) are usually made in the declarative mood, future tense, first person, even though they are a species of self-imperative. However, they rarely make use of the imperative mood (with the exception of some self-addressed comments, such as: "Get a grip, Schiffrin").

9.2.4.4 A State-Trace Approach to Requestive Speech Acts

I have outlined the way that I intend to treat questions in the discussion so far; now, in Table 9.7 and the following analysis, I will show how they can be accommodated into the state-trace model without further amendment to the proposed structure.

State-Trace Model	*Type of Question*	*Example*
REQUEST-ASSERT (≈ PRE-ASSERT)	*WH-*	'Where is that?'
	Alternative	'Is that in Sheffield or Leeds?'
REQUEST-AGREE (≈ ASSERT)	*Yes/No* (1)	'Is that in Leeds?'
	(2)	'Isn't that in Leeds?'
	Tag (1)	'That's in Leeds, is it?' ≈ *Yes/No* (1)
	(2)	'That's in Leeds, isn't it?'
	(3)	'That isn't in Leeds, is it?' ≈ *Tag* (2)
	Declarative	'That's in Leeds?'
REQUEST-DIRECT (≈ PRE-DIRECT)	*WH-*	'Where am I taking it?'
	Alternative	'Am I taking it to Sheffield or Leeds?'
REQUEST-ACCEPT (≈ SELF-COMMIT)	*Yes/No* (1)	'Am I taking it to Leeds?'
	(2)	'Am I not taking it to Leeds?'
	Tag (1)	'I'm taking it to Leeds, am I?' ≈ *Yes/No* (1)
	(2)	'I'm taking it to Leeds, aren't I?'
	(3)	'I'm not taking it to Leeds, am I?' ≈ *Tag* (2)
	Declarative	'I'm taking it to Leeds?'
REQUEST-SELF-COMMIT (≈ PRE-SELF-COMMIT)	*WH-*	'Where are you taking it?'
	Alternative	'Are you taking it to Sheffield or Leeds?'
REQUEST-ACCEDE (≈ DIRECT)	*Yes/No* (1)	'Are you taking it to Leeds?'
	(2)	'Aren't you taking it to Leeds?'
	Tag (1)	'You're taking it to Leeds, are you?' ≈ *Yes/No* (1)
	(2)	'You're taking it to Leeds, aren't you?'
	(3)	'You're not taking it to Leeds, are you?' ≈ *Tag* (2)
	Declarative	'You're taking it to Leeds?'

Table 9.7 A state-trace approach to interrogatives.

In the first column of Table 9.7, I specify the name of the speech act that the question performs in the state-trace model, along with the name of the descriptive or prescriptive speech act to which it is equivalent (shown in brackets and by the symbol '≈'). As can be seen, all questions are interpreted either as a prompt for an initiating act: a REQUEST-<*focus-setter*> (a kind of PRE-act); or as standing in place of the initiating act itself: a REQUEST-<*response*>. All questions that are answered with a whole proposition, even if this is in an elliptical form (such as 'To Leeds' for instance after the question 'Where are you going?'), are of the type REQUEST-<*focus-setter*>: i.e. REQUEST-ASSERT, REQUEST-DIRECT, and REQUEST-SELF-COMMIT. *WH-* and *Alternative* questions are of this type, which can typically be identified because they cannot be 'Agreed' with. All questions that can be answered with 'Yes' or 'No' are of the type REQUEST-<*response*>: i.e. REQUEST-AGREE, REQUEST-ACCEPT, and REQUEST-ACCEDE. *Yes/No*, *Tag* and *Declarative* questions are of this type, and they can be 'Agreed' with (although some

formulations of these types of question are more appropriately answered in this way than others).

So, I have separated out different kinds of questions according to whether the speaker is ambivalent or committed to the content. This seems appropriate, as although in the example 'Where am I taking it?' the speaker is committing himself to the act of taking something somewhere, because the proposition is not complete, the commitment is dependent on the answer. Sometimes this is a difficult distinction to make, and I suspect that the types of questions may not fall into these separate categories in quite as neat a manner as Table 9.7 suggests. A more thorough investigation would be needed to ensure that the model defined here is comprehensive.

It could be argued that REQUEST-<*response*> speech acts would be more properly labelled REQUEST-AGREE/DISAGREE, REQUEST-ACCEPT/DECLINE, or REQUEST-ACCEDE/REFUSE respectively, as they are equally well answered by the negative speech acts as by the positive. I choose to label these types of request by the positive response alone not only for the sake of brevity, but also because no one ever requests to be disagreed with, or to be declined or to be refused. However, just because the negative response is the dispreferred answer, this does not preclude its use. As can be seen from the example below, a REQUEST-ACCEPT can just as well be answered by DECLINE as by ACCEPT.

(1) **Terry**:	shall I say something at this point?	REQUEST-ACCEPT
(2a) **Kevin**:	I- I- yes, I'd love you to say something.	ACCEPT (1)
(2b) **Kevin**:	no thanks, not just at the moment.	DECLINE (1)

This distinction between a REQUEST-<*focus-setter*> and a REQUEST-<*response*> is an important one as it differentiates between those questions that have no commitment towards the proposition and those that have some, albeit faint, commitment one way or another. I will return to this point and highlight it with examples in the following discussion. If this can be shown to be a sound classification of questions, then it would be a very elegant method of incorporating interrogatives into the state-trace structure, as the speech acts that follow on from questions would use the same schemata developed in the previous sections.

In the second column of Table 9.7, I list the question types that are interpreted as corresponding to each speech act, and in the third column I give examples of these question types. In some cases the class of requestive speech act to which a type of question should belong is not altogether clear-cut. Let us look at each type of question in turn in order to get a better idea of how they fit into the model.

WH-questions: *WH*-questions are always of the type REQUEST-<*focus-setter*>, because the speaker is expressing no commitment towards a specific answer. Usually, *WH*-questions are seen as requests for missing information. While this is true at the content level of such questions, at the discourse function level they are actually used as incomplete requests not only for the information supplying assertive acts, but also for directive and commissive acts. This is because the answer expects not only the provision of the completed content, but thereby also the provision of the completed assertive, directive or commissive act. E.g.

(1)	**A**:	Where am I taking it?	REQUEST-DIRECT
(2)	**B**:	Take it to…	DIRECT(1)[21]

This is as far as I am aware the only attempt to fit *WH*-questions into a more generalised framework. Tsui (1994: 73) notices that not all kinds of *WH*-questions behave in the same way. She points out that this can be seen in the inappropriate use of 'Thanks' in the contrasted examples shown below:

A:	What's the time?		**A**:	What time shall we meet?
B:	Seven.		**B**:	Seven
A:	Thanks	*	**A**:	Thanks

However, Tsui does not draw the conclusion that it is because the first *WH*-question has a descriptive content, while the second has a prescriptive one that this discrepancy of usage occurs.

Alternative questions: Like *WH*-questions, *Alternative* questions are also always of the type REQUEST-<*focus-setter*>.

(1)	**A**:	Is it raining or snowing?	REQUEST-ASSERT
(2)	**B**:	Snowing	ASSERT (1)

Alternative questions can be formed syntactically in two different ways and these are very strongly related to both *WH*-questions and *Yes/No* questions. They are like *WH*-questions because they expect the answer to be resolved from a set of (circumscribed) options. In the case of *WH*-questions the set is delimited by the property-type of the missing information, whereas in

---

[21] Note that the speech act DIRECT is referential in nature when used in answer to a question. The preconditions in Tables 9.1 to 9.3 for the recognition of the ASSERT, DIRECT and SELF-COMMIT speech acts will have to be updated appropriately for the case when they follow a PRE-focus-setter question. It is this that will distinguish an answer from an unprompted directive.

the case of *Alternative* questions the set is delimited by a specified list of alternative propositions (or property-values). Note their similarity in the examples below (sentence taken from Tsui 1994:74-76 – see reference for her discussion of these types of question):

WH-: 'Which ice-cream would you like?'
Alternative: 'Which ice-cream would you like: chocolate, vanilla or strawberry?'

Alternative questions are like *Yes/No* questions as well because instead of the set of answers comprising either the proposition or its negation, the answer is a set of alternative propositions. Compare the following two questions:

Yes/No: 'Would you like an ice-cream?'
Alternative: 'Would you like a chocolate, vanilla or strawberry ice-cream?'

Both of the formulations for *Alternative* questions shown in the examples above carry an assumption that the addressee does want an ice-cream in the first place and that the only point under discussion is the flavour that the addressee wants. If the speaker is not sure whether this assumption can be made, he will often ask the addressee the *Yes/No* question 'Would you like an ice-cream?' first before listing the alternatives. Otherwise, the addressee can always deny the presupposition that he would indeed like an ice-cream.

Yes/No questions: *Yes/No* questions are always of the type REQUEST-<*response*>, which can be composed in two ways, positively or negatively as exemplified in Table 9.7. When formulated positively as shown by the example (1), they are almost entirely neutral with respect to the speaker's commitment to the proposition. By using a *Yes/No* (1) question, the speaker displays next to no prior assumption about what the answer should be. However, I propose that simply by forming the proposition in one way rather than another, the speaker betrays enough of a bias in favour of the propositional content for this not to be interpreted as a REQUEST-<*focus-setter*> act. Admittedly this is a slightly contentious position. It could be argued that in fact *Yes/No* (1) questions should not be included in the REQUEST-<*response*> category, but interpreted more like *WH-* and *Alternative* questions. *Yes/No* (1) questions that are composed using the modals 'would' and 'could' are more clearly in the REQUEST-<*response*> category than those that are not, e.g. 'Would you pass the salt please?'. If I reply 'Yes' to this question, then this would clearly be interpreted as an ACCEDE rather than a SELF-COMMIT act.

Yes/No (2) questions, formulated in the negative, definitely do have a biased commitment towards the negated proposition. If we compare and contrast between 'Is that in Leeds?' and 'Isn't that in Leeds?', we intuitively feel that the answer 'Yes' is more strongly presupposed in the latter than in the former. Here the negative marker is not used to form the negative proposition 'That isn't in Leeds' but to introduce an element of doubt to the proposition: 'That

is in Leeds', so that the addressee is asked to confirm that the proposition is true. A similar effect occurs with *Tag* questions to which I shall turn shortly. It is the addition of the marker expressing doubt (in this case the adverb 'not') that causes these to be interpreted with a stronger degree of commitment to the proposition than the *Yes/No* (1) formulation. In fact, any question (i.e. *Yes/No*, *Tag*, or *Declarative*), which includes a marker indicating doubt about the proposition, will be interpreted as increasing the speaker's commitment towards one answer rather than another. In the example we have discussed so far, it is the use of the adverbial marker 'not' that introduces the element of doubt. Other adverbs will also have the same effect:

'Is that **really** in Leeds?'
'Has the train left **already**?'

It will be easier to distinguish such cases with the addition of the concept of degrees of commitment as discussed in Section 10.1.3.

Tag questions: *Tag* questions are always of the type REQUEST-<*response*>, and can be composed in three ways as exemplified in Table 9.7. They are very similar to *Declarative* questions (discussed below), because they are composed of the *Declarative* formulation plus the *Tag* question at the end. Participants will often append the tag as an afterthought in order to clarify that what might be taken as a straight ASSERT is in fact a request for confirmation.

Tag (1) questions are a simple reformulation of *Yes/No* (1) questions, e.g.: 'Is that in Leeds?' and 'That's in Leeds, is it?'. These types are equivalent and should be treated in the same way.

Tag (2) and (3) questions are very strongly related to each other. However, let us look at the difference between these two different forms of *Tag* questions. Both are biased towards a positive answer (because they contain an embedded assertion), but the second is even more so than the first. Consider:

A: That's in Leeds, isn't it?
B: No.

The answer 'No' here signifies the disagreement 'That is not in Leeds', or 'Your assumption is incorrect'. Compare this with:

A: That isn't in Leeds, is it?
B: No.

In this case the answer 'No' stands for the agreement 'That is not in Leeds', or 'Your assumption is correct. One would perhaps have expected 'No' to be a disagreement as in the prior example, because the proposition is the negated one 'That isn't in Leeds'. The answer

'Yes' to this question is completely ambiguous, but is most likely also to stand for an agreement. Answering a *Tag* (3) question, which contains a negated proposition, is problematic and would probably require further clarification in either case as shown in Table 9.8:

REQUEST-AGREE (ASSERT)	'That's in Leeds, isn't it? ('That's in Leeds')	'That isn't in Leeds, is it?' ('That isn't in Leeds')
Default Interpretation of 'Yes' or 'No' Responses	'Yes' = AGREE 'No' = DISAGREE	'Yes' = AGREE 'No' = AGREE
Interpretation of Clarified 'Yes' or 'No' Responses	'Yes, it is' = AGREE 'Yes, it isn't' = * 'No, it is' = * 'No, it isn't' = DISAGREE	'Yes, it is' = DISAGREE 'Yes, it isn't' = AGREE 'No, it is' = DISAGREE 'No, it isn't' = AGREE

Table 9.8 'Yes' and 'No' responses after a negated content focus-setter.

It is probably for reasons of scope that participants in a conversation can be confused in the case of the *Tag* (3) question (with the negated proposition) by the bare answer of 'Yes' or 'No' alone. In other words, it is ambiguous whether the answer refers to the assumption or to the question. Oddly, for *Tag* (3) questions both 'Yes' and 'No' might be taken as acts of agreement. Perhaps this is because there is a greater sense that the speaker is committed to the truth of the negated proposition and does not expect to be gainsaid.

It is interesting to note that exactly the same phenomenon is encountered when considering the response to descriptive and prescriptive focus-setters with negated contents (see Table 9.8 above). This is encouraging as it backs up the claim of equivalence: REQUEST-AGREE ≈ ASSERT.

Another way of dealing with the *Tag* (3) question might be to re-label these kinds of speech act as REQUEST-DISAGREE, REQUEST-DECLINE or REQUEST-REFUSE respectively. In these cases the speaker is after all prospecting for a negative response. However, I think these labels would be misleading because as we have mentioned previously, although they are carried out using a negated proposition, the types of act expected in reply are still AGREE, ACCEDE or ACCEPT.

Although I have listed the two options *Tag* (2) and (3) questions (both the straightforward and the negated proposition) as equivalent in Table 9.7, as I have shown in the discussion so far, in conversation they are dealt with differently. At the functional level however, they are indeed equivalent. The difference might be represented at a later date by different levels of commitment to the proposition in an extension of the state-trace model, as described in Section 10.1.3.

Declarative questions: *Declarative* questions are always of the type REQUEST-<*response*>. This is because commonly the speaker has a commitment to the content of the question and is simply looking for confirmation from the addressee:

'You prefer that one?'

Plainly there is a relationship between the *Declarative* question and a normal ASSERT act which is also formulated using the declarative mood and which would be lexically and grammatically identical; the only means of distinguishing the two is by intonation.

Often a *Declarative* question will be performed to confirm a previous utterance by a speaker, and may additionally indicate surprise or disbelief:

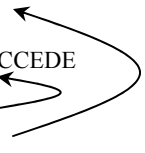
A: I prefer that one.
B: You prefer that one?

Perhaps therefore there is a case to be made for classifying *Declarative* questions separately from other types of questions? After all, they are rarely performed in isolation, when the context space is entirely empty.

9.2.4.5 Some Difficulties with the Requestive Representation

One would like to extend the list of requestives given in Table 9.7 to include requests for other speech acts, such as REQUEST-REPEAT, REQUEST-RETRACT, etc. However, this is not possible at this level, because there comes a point when the speech act being requested is in fact the action that is prescribed for the agent by the speaker. To exemplify this problem, let us look at the case for REQUEST-REPEAT.

(1)	**B**:	The bus will leave at eleven.	ASSERT
(2)	**A**:	Could you repeat that please?	REQUEST-ACCEDE
(3)	**B**:	Yes,	ACCEDE (2)
		I said…	REPEAT (1)



This highlights two problems. Firstly, the use of the label 'ACCEDE' to indicate a positive response to a directive (or request-commissive) misleads us into thinking that the response itself is the act of accession, which it is not. It is only a verbal commitment to a future act of accession. In the example above, it is the REPEAT act that is the real act of accession. This is true for all prescriptive acts (except that the action required need not be the performance of a speech act as it is in this case).

Secondly, we would wish somehow to connect the act of repetition not only to the assertion being repeated, but also to the implicit prescription '**B**: ▸ REPEAT (1)' within the REQUEST-ACCEDE. The REPEAT (1) act will have the consequence of closing the context space opened by the REQUEST-ACCEDE by completing the asked for prescription. I cannot currently do this with the model as it stands, because I have not integrated it into a wider theory of action. The contents of the prescription are not analysed, and so we simply treat the verbal act of accession as the closure of the REQUEST-ACCEDE act. Again this problem extends to all prescriptions.

If I am not careful at this point, then the same criticism that I levelled at the VERBMOBIL scheme for requestives will, with equal justice, be levelled at the state-trace model: that the inclusion in the label for the speech act of the type of action requested, would soon mean that distinct speech acts would proliferate uncontrollably (e.g. REQUEST-OPEN, REQUEST-INTRODUCE, etc.). It is important to emphasise that the list of different acts of REQUEST is circumscribed in nature and ranges over basic descriptive and prescriptive acts; they either prompt for a focus-setter, or stand in place of one.

There are two additional issues that are closely related to the problem of the distinction between a request for the performance of a speech act and a request for any other action as described above. Firstly there is often no need for an explicit verbal accession to perform the action prescribed in the requestive. The prescribed action can be acceded to without recourse to words at all (unless of course it is a speech act that is prescribed). In the example given for the case of REQUEST-REPEAT above, **B**'s answer in (3) begins 'Yes', which is interpreted as a commitment to ACCEDE. This might just as easily have been omitted.

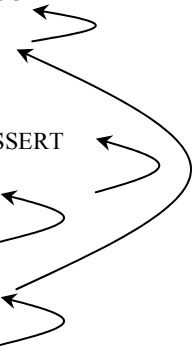
Secondly, there are a number of ways of performing what everyone would recognise as a paraphrase of what is essentially the same act; yet each of these would be interpreted in the current model as performing subtly different acts. E.g.:

'What is the time?'	REQUEST-ASSERT
'Could you tell me the time please?'	REQUEST-ACCEDE
'Tell me the time'	DIRECT

Apart from considerations of politeness, can we really claim that there are significant differences between a direct REQUEST-ASSERT act, and an indirect REQUEST-ACCEDE or a DIRECT act to elicit an ASSERT? All of these speech acts can be answered in exactly the same way: '[Yes,] It's half past twelve'. Just as in the case of REQUEST-REPEAT, an accession to the acts REQUEST-ACCEDE or DIRECT could be shortcut by the performance of the act of telling the time as requested.

There are further problems with the way that speech acts are identified. We cannot ignore the content and form of an utterance when trying to determine which speech act it may be an instance of. There are some ways of forming requestive (and prescriptive) utterances that do not follow the pattern that is expected of them. Take the example conversation:

(1)	**Mandy**:	do you want another drink?	REQUEST-ACCEPT
(2)	**Julie**:	no I'll wait-	DECLINE (1)
(3)		oh it's twelve o'clock.	ASSERT
(4)		what time do you want to eat?	REQUEST-ASSERT
(5)	**Mandy**:	'bout one o'clock?	ASSERT (4)
(6)	**Julie**:	yeah	AGREE (5)
(7)		no I'll wait till then.	REPEAT (2)
(8)	**Mandy**:	okay.	DEFER (7)



The utterance in (1) is labelled REQUEST-ACCEPT because intuitively we feel it is an offer by Mandy to make Julie another drink. However, we could equally well interpret (1) as a REQUEST-AGREE because the surface form of the utterance can be heard as the expression of a wish to be informed whether 'Julie wants another drink' is true or false. Clearly this is not all that Mandy wants to know in performing this speech act. If Julie answered 'Yes please' and Mandy then did nothing about satisfying Julie's want, there would be something very wrong with the conversation. We know that this is indeed a REQUEST-ACCEPT act because the utterance in (1) actually stands for 'Shall I make you another drink?'. This interpretation is partly dictated by the oddness of verbs such as 'want' and 'like', which express a participant's preferences and so in effect reverse the direction of the prescription by a procedure of inferring that the participant wants to do something about those preferences. E.g.:

'I'd like an ice-cream please'	DIRECT/REQUEST-ACCEDE
'Would you like an ice-cream?'	SELF-COMMIT/REQUEST-ACCEPT

Without some method of inferring one meaning from another, without the semantic knowledge of how these kinds of verbs work, we are unable to account for such utterances simply. This is a limitation of the state-trace model and will need considerable further research.

In this Section 9.2.4, I have tried to show how questions might be incorporated within the model as it currently stands. It must be admitted that there is some way to go before one could be truly satisfied that this integration has been carried out entirely satisfactorily. The study presented here is far from complete or rigorous; much further work would be required, particularly gathering evidence from real data, to be sure that the conclusions drawn are accurate. However, having said this, I believe that I have made a good start.

So, in order to begin to deal with some of the issues that remain unsupported or unclear in the classification of what I have labelled 'requestive' speech acts, the following questions will need to be addressed:

- What kind of speech act are questions?
- How many different types of interrogative speech acts are there?
- How do questions adjust the focus of conversation?
- Do questions have complex interactions with the other speech acts in the model?
- How are questions related to other speech acts?

Before closing this chapter with a detailed worked example of a real piece of conversation, I will very briefly turn my attention to some speech acts that we have not accounted for so far. This will highlight some potential holes in the model as it currently stands, and suggest a means for dealing with them.

9.2.5 Other Speech Acts

The discussion of assertive, directive and commissive speech act types in Sections 9.2.2 and 9.2.3 may have inadvertently given the impression that the two types of direction of fit are not inter-related and act independently of each other. However, there are speech acts, particularly descriptive ones, that will have considerable effects upon the state of activity of other speech act types. So, is the direction of fit idea too simplistic? Do we need a broader theory to encompass both types of direction of fit?

There are (at least) two other ways of responding to prescriptive and requestive speech acts that are not covered in the model so far. These are acts of descriptive type that have an effect on the state of activity of the current context space because their propositional contents match a prior prescription or request in some well-definable manner.

First is the case when the responder defeases the initiator's prescription or request by asserting that the state of affairs that is being prescribed or requested to be brought about is already true in the current context (and therefore the action cannot be carried out by the agent, who is commonly the responder). For example:

(1) **A**:	Shut the door	***B***: ▸ *The door is shut*	DIRECT
(2) **B**:	It's already shut	*The door is shut*	DEFEASE (1)

We could view this as a kind of REFUSE act, with the 'No' implicit in the explanation for the refusal. However, one does not strictly feel that **B** is refusing, simply informing **A** why he is unable to comply with the direction, and discharging himself from the obligation of having to commit himself one way or another.

The second use of an assertive that might have an effect on the state of activity of a context space opened by a prescription or request is very similar to the previous. Often when a participant has completed a prescribed or requested action, he will REPORT that the propositional content is now true in the context.

(1)	**Mandy**:	oh, get me a pint of milk while you're out.	DIRECT
(2)	**Julie**:	'kay	ACCEDE (1)
…			
(3)	**Julie**:	I got that milk for you.	REPORT (2)
(4)	**Mat**:	No you didn't, I did.	DISAGREE (3)

Note that the completed action is reported in the past tense. Do we want to call any type of utterance that reports a past event, or action, a REPORT act (e.g. 'I went out with Anna last night')? What is the difference between this and an ASSERT act? Is it only a REPORT if the content relates to a prior prescription or request? More work would be needed to assess whether these should indeed be included in an analysis of basic speech acts as defined in this chapter.

What we can say is that DEFEASE and REPORT are variants of the ASSERT act. Unlike ASSERT they are not focus-setters and are therefore referential in nature, but they are responded to by the same speech acts as ASSERT, as the example above shows.

I include a brief discussion of these acts because they would have the effect of closing a previous prescriptive context space and are therefore of significance to the state-trace model. The inclusion of such patterns of acts may considerably complicate the model, so I do not propose to go into further detail here. Further work will be required at some point to unify the definitions of the different types in a more coherent manner, and expand the list of preconditions in the context space for the identification of each speech act.

Having spent some time describing the development of the model, and the breakdown of how speech acts are characterised within particular context spaces, I now show how the model might work on the example problem piece of conversation given in Chapter 1.

9.2.6 Applying the Model to Example Conversation 1.1

At the beginning of this dissertation, I highlighted some of the grave difficulties in analysing spoken conversation with reference to an example conversation taken from the British National Corpus (see Conversation 1.1 in Chapter 1 for details). I suggested a provisional analysis of the function of the various utterances in a sub-section or conversational 'thread' of Conversation 1.1 in Figures 1.1 and 1.2, producing a form of 'benchmark' against which to test any functional model of speech acts. To the best of my knowledge, there are at present neither any annotation

schemes nor any structural models that will fully explain all the kinds of communicative behaviour shown up in this example. What I wish to do now is compare this with the analysis that would be given by my model and account for the differences where they occur.

Before doing so however, I wish to look briefly again at solipsism and how this applies to the model as I have defined it. Throughout this dissertation, and particularly in Chapter 8, I discussed the importance of a solipsistic approach to modelling language. I wanted to show that one has to allow for the ability of different people's mental models to come apart and end up in different states. However, the model that I have proposed here seems to be a 'one size fits all' model, much like its theoretical predecessors. In describing its development and when walking through examples, it has been necessary to apply the model from one point of view only, primarily to explicate how it works. In this sense, the analysis of Conversation 1.1 given in this section is solipsistic in nature because it represents my perception of the conversation as an outsider, but it is not solipsistic in that I do not present either June or Albert's models of the same conversation. That is because I am neither June nor Albert. I am like a silent third participant of the conversation, who listens without taking part and has his or her own interpretation of what goes on. I cannot ever be sure that my model is the same as that of the other participants. For the most part, unless I specifically say otherwise, I assume that they are the same; this is an assumption that could be quite incorrect. This in part is a side effect of analysing transcribed conversation, and indeed, of only ever really being able to present my own version of reality, as much as I can perceive it, or am capable of conveying it.

The model itself is intended to be a solipsistic one in that it is designed to be the representation of one agent's context space. If I were to try to implement the model say as a program for autonomous agents who had to communicate to each other via an evanescent medium such as speech, then even though the respective interpretive programs might be identical for both agents, the information stored at the end of the interaction might not be the same. Miscommunication might occur because one agent has extra information, which allows it to make inferences that the other is unable to, or because the information is passed in such a way that the interpretation is ambiguous, or because information is lost during transference due to mishearing or noise, etc[22]. As it is the semantic representation of each utterance in the conversation that informs the model, and as the 'output' speech act interpretation of the model will be further interpreted (as I hope to show in Chapter 10), it is at that point that inconsistencies might be introduced.

---

[22] Note that, unlike human beings, autonomous computer agents will never think that they have said something when in fact they have not or vice versa. Forgetfulness, unless built in on purpose to simulate human cognitive weakness, is purely an animal phenomenon.

What I am claiming is that the underlying human interpretive model is generic (to speakers of the same language at least), but how the utterance is 'taken' will depend on many other factors, such as personality, intelligence, background knowledge, privileged information, emotional state and accuracy of hearing ability, to mention a few. These factors are not entirely predictable or controllable, but the underpinning functional structure is. This is very difficult to show diagrammatically or with examples, as we are not generally able to see inside other people's heads. There is also a big difference between the interpretation of 'live' conversation, as it is actually occurring, and that of transcribed examples. I have tried to characterise some of the mechanisms we use for speech act recognition, with the thought that at some point in the future it might be possible to simulate the kinds of conditions under which errors are made and hence show how it is that mental models come apart.

Having clarified this point, which will become relevant again during the course of the following discussion, let us now try to apply the model to Conversation 1.1 and critically analyse the results. There are eight utterances in this specific context space; let us go through each utterance one at a time.

(1) **June**: Shut door.

This is a straightforward directive issued in the imperative mood (with a slightly ungrammatical contracted 'the'). As there are only two participants in the conversation (that we know of), the agent is obviously intended to be the person who is not the current speaker and the only other person in the conversation, Albert. We interpret this then as "Albert, shut the door", or "Albert, bring it about that: the proposition 'the door is shut' = true". This is added to the current context space (in the following representation, X stands for 'the door is shut', the symbol ▸ stands for 'bring it about that' and the symbol ⇒ stands for 'is interpreted as'):

CS1	**June:**	**Albert:**
	(1) Albert: ▸ X ⇒ DIRECT	∅

We interpret Albert's next utterance as an act of refusal precisely because the context space is no longer empty. It is because of June's previous utterance that Albert can leave the proposition underspecified; "I can't" is the contracted version of "I can't shut the door".

(2) **Albert**: I can't

Note that although we say that Albert is refusing – which is not an inaccurate interpretation as it happens – in leaving the analysis at that there is an element of meaning lost; "I can't" is not the same thing as "I won't". The former implies that there is a good reason why the speaker is unable to comply with the prescription, while the latter is just a straight, uncompromising refusal. Our updated context space looks like this:

CS2	**June:**	**Albert:**
	(1) Albert: ◗ X ⇒ DIRECT	(2) Albert: ¬ ◗ X ⇒ REFUSE (1)

In utterance (3), Albert starts to explain why he cannot comply, but never finishes. At least, that is the impression we have according to our conversational expectations and because of the positioning of the utterance immediately after the refusal. We cannot know this for an absolute fact – Albert could just as easily have been intending to say something entirely different. There is also an element of post-processing here as we know that Albert explains himself later, we assume that he begins to do so here.

(3) **Albert**: I'm going to er…

In the analysis presented in Chapter 1, this act is labelled an incompleted justification. Perhaps it might have been equally valid to leave this act without any interpretation (as, after all, the speaker does not finish what he is saying). In my model I have not allowed for the act JUSTIFY. The most one could say with my model as it stands is that it is the beginning of a SELF-COMMIT. With the addition of some ability to make inferences, we might be able to piece together that it is the beginning of a SELF-COMMIT which will imply 'Albert: ¬ ◗ X', therefore making this an act of repetition of the previous refusal, but we are not yet able to encompass justifications and explanations with the model. I shall return to this point again shortly. I discuss more fully how the model might be extended in the future to deal with this phenomenon in Section 10.1.6.2 in the last chapter.

So, the context space, according to our two possible treatments above of this utterance, will either remain unchanged (because the assertion might be entirely unconnected with the current context space) or be added as a repetition (because we have not yet accounted for acts of explanation).

CS3	**June:**	**Albert:**
	(1) Albert: ◗ X ⇒ DIRECT	(2) Albert: ¬ ◗ X ⇒ REFUSE (1) [(3) Albert: ¬ ◗ X ⇒ REPEAT (2)]

Either way, the subsequent interpretations will not be affected. After the first three utterances, we have a small lapse of time when other conversations with their respective context spaces are pursued. The context space remains 'open' because there is a conflict between the participants' stances to the proposition. When June resumes the old context space again, she cannot now use an ambiguous shortened form because the context space has become superseded by others. She brings the context space back into focus by reissuing the direction in full.

(4) **June**: Shut the door.

This is not merely an act of repetition, but one of insistence on her original stance. We interpret this as INSIST precisely because of the presence of Albert's negative commitment in the context space, which importantly comes <u>after</u> the original directive. The context space is updated accordingly

CS4

June:	**Albert:**
(4) Albert: ◂ X ⇒ INSIST (1)	(2) Albert: ¬ ◂ X ⇒ REFUSE (1) [(3) Albert: ¬ ◂ X ⇒ REPEAT (2)]

Albert then insists on his inability to comply.

(5) **Albert**: I can't shut door. <*pause*>

Leaving the context space looking like this:

CS5

June:	**Albert:**
(4) Albert: ◂ X ⇒ INSIST (1)	(5) Albert: ¬ ◂ X ⇒ INSIST (2)

Both participants have now reached a total impasse. Their respective stances are contradictory to the maximum level. Either one participant must give up their position, or the directive must remain unfulfilled. It is at this point that June gives way to Albert's refusal by offering an alternative.

(6) **June**: I will.

This is the contracted form of "I will shut the door". In one sense this is a straightforward self-commitment. But it is also an act of deferral due to the inference that if June closes the door, then Albert will not close the door. This requires the addition of some transformation rules to the model to show how one thing can stand for another without being contradictory. So, it is not the utterance itself that is the deferral, but the inference of the utterance. The updated context space will now resemble the following:

CS6

June:	**Albert:**
(6) June: ◂ X ⇒ SELF-COMMIT → (6a) Albert: ¬ ◂ X ⇒ DEFER (5)	(5) Albert: ¬ ◂ X ⇒ INSIST (2)

Note here that something rather odd happens. We do not know whether Albert in fact hears June's SELF-COMMIT act or not as he makes no response. We know from the context that as he is in the process of making a cup of tea, it is likely that the conversation is taking place between two different rooms, possibly with raised voices. I imagine June sitting in the living room with the tape recorder while Albert is in the kitchen next door with the door open. Perhaps June's "I will" was half spoken under her breath? If this picture is accurate, then there is a possibility that Albert might have missed this act altogether and that at this point his context space has come

apart from June's; he could be one update behind her. This would account for the lack of feedback and for the fact that it is then Albert after all who completes the action despite June's self-commitment. This is just an idea; there is no way of really knowing.

Although I have updated the context space above so that June no longer has a commitment that Albert close the door (because of the inferential deferral), there is no doubt that her wish to have the door closed still persists in the context. Whether that is because Albert takes her goal upon himself through norms of co-operative behaviour, or whether he has not heard June's self-commitment, is largely irrelevant except for repercussions on the storage of commitments in the context space representation and the subsequent speech act interpretations because of the referential nature of speech acts. Throughout this conversation it is an accepted fact that June wishes the door to be closed. It is how this action is brought about that comes under discussion. Perhaps the context space itself is identified by this underlying proposition – that it is June's desire that the world should be changed such that 'The door is shut' becomes true.

The next utterance highlights some of the weaknesses of my model as it stands so far.

(7) **Albert**: I've shut the door now.

If Albert has not heard June's last utterance then the functional explanation for his utterance is fairly easy to account for and his context space will be as follows (the symbol ◂ stands for 'brought it about that' – a completed action):

CS6′

June:	**Albert:**
(4) Albert: ▸ X ⇒ INSIST (1)	(6) Albert: ◂ X ⇒ REPORT → (6a) Albert: ◂ X ⇒ DEFER (4)

The problem here is that this is an act of reporting an act of deferral (or belated accession). The deferral itself is the physical act of closing the door, or the intention to close the door. The fact that it has already happened is announced here by the past tense of the main verb. How would this be shown in the model? How does this interpretation come about if Albert has indeed missed June's last utterance? Does June infer that this is so and so use the above context space for interpreting Albert's utterance? Or is something else occuring here?

CS7

June:	**Albert:**
(6) June: ▸ X ⇒ SELF-COMMIT → (6a) Albert: ¬ ▸ X ⇒ DEFER (5)	(7) Albert: ◂ X ⇒ REPORT → (7a) Albert: ◂ X ⇒ DEFER (4)

If *CS7* (which comprises all moves so far) is the correct context space for both participants, then Albert's act of closing the door is actually in direct conflict with June's self-commitment to shut the door. His action might even be seen as defiance in certain contexts. It is because of the

persistence of June's original prescription that Albert's utterance can be correctly interpreted. The important thing is that X is resolved.

The way I have dealt with this here is to add a marker to keep track of whether the actions are completed or not, according to the tense of the main verb. Ideally we would also want to integrate actions taking place in the 'world' (virtual or otherwise) with the speech actions. Clearly the model in its current form is inadequate for dealing with this feature of reported compliance.

Finally, let us look at Albert's last utterance.

(8) **Albert**: I've finished running about.

This is arguably the most complicated utterance of all in the conversation. At present the model would see this as a REPORT of a completed action or as an ASSERT; on the surface there appears to be no cohesion with the current context space at all. Why is it that this utterance has any bearing on the door conversation? Coming immediately after the act of deferral however, it becomes obvious that this is an explanation for Albert's earlier non-compliance. The inferences required to come to this conclusion are far from trivial. One would have to know that 'running about' might entail going through open doors, which is why he was unable to shut the door before and why he is able to do so now. I admit that my model makes no provision for these kinds of inferences. This would be seen as a plain assertion of fact – not in itself wrong, but certainly not functionally complete. Currently there is no way of showing that there is any connection between this utterance and the previous. In Section 10.1.6.2, I try to show how the model might be extended in order to cope with justifications and explanations – I do not underestimate the difficulties inherent in doing so.

The assertive would pointlessly open a new context space (here Y stands for 'Albert has finished running about'):

CS8	**June:**	**Albert:**
	∅	(8) Albert: ◂ Y ⇒ REPORT [(8) ASSERT (Y)]

However, whether we recognise this last utterance as an explanation or not, we now have our context space resolved – strangely both according to, as well as against, the express prescriptions of the initiating participant (which is partly what makes me think that Albert might not have heard June's last utterance).

We can see now that my model reproduces a good approximation to the functional description of the flow of Conversation 1.1, which we defined as our ideal in Figures 1.1 and 1.2 of Chapter 1. I cannot say with complete certainty whether this would be so for all conversations. My

research has to date all been carried out through painstaking hand analysis of various conversations – both available sections of spoken corpora and my own recordings of the conversations of friends. As I stated in Chapter 1, the results of this thesis as a consequence are almost entirely qualitative in nature. I have taken this preliminary model as far as I am able to.

In the final chapter, I shall look in more detail at the various limitations of the model as it stands and suggest ways in which these could be addressed to extend the model in the future and encompass a greater range of phenomena. I shall end by considering, in the light of this discussion, what I have achieved with the development of the model so far.

Chapter 10
Conclusion

I bring not only all I wrought
Into the faltering words of speech,
I dedicate the song I sought,
Yet could not reach.
(W. Somerset Maugham)

Natural Language Processing has come a long way since its humble beginnings in text processing forty years or so ago. There are now systems that will understand voice commands and are able to interact with users directly through the medium of speech. To give concrete examples, take the voice activated Windows system being developed using Microsoft MindNet technology (Yale 1998), British Telecommunication's spoken email management system (Downey et al. 1998), a cinema programmes information line in Germany (Sympalog), a London restaurant guide (Philips), a stock market quotes line in English, French and German (Speechworks), as well as the various computer-operated rail travel information systems already successfully running in places like the Germany (DASA and Philips) and Sweden (Philips).

This progress is all well and good, but we are still very far from being able to converse in any natural way with our machines (see McTear, in press, for a comprehensive overview of the development of modern task-oriented spoken dialogue systems). However sophisticated NLP programs have become, they are still hampered by a number of related problems. Most systems are still very much domain dependent, and are unable to deal with anything that falls beyond the scope of their function. I cannot ring up the German train information system and ask the following (obviously, in German):

'Can you tell me whether the line's still flooded near Munich please?'

This would be a natural question to ask if the user intended to travel to Munich and needed to know whether he could; on the surface there is apparent incoherence, but by inferring the intention, a human operator would have no problems answering this question; even if he could not infer the user's intention, a human being would be able to jump out of the task-domain strictures and answer the question. People ask for unexpected information more often than was originally thought, thus bringing down the number of 'satisfied customers' who use such information lines.

Another problem is that there is a heavy emphasis on statistical methods for determining a speaker's meaning. These techniques work well on 'clean' speech, but in every day life speech is subject to background noise and, in the case of telephonic communication, to distortion and

loss of detail through transmission. It is the ability to reason about what a person is saying that will give an accurate interpretation of an utterance.

While it is not possible to represent computationally at this time all the knowledge necessary for such rationalising, the state-trace model (STM) I have defined in Chapter 9 goes some way towards providing a framework for such capabilities by describing a grammar of discourse potentially able to distinguish between a large number of distinct speech acts. This is more than other models presently are able to do, and is in itself an original contribution to computational linguistics. In the next section I will outline some of the deficiencies of the STM and how these might be addressed, and the model improved and extended.

10.1 Limitations and Future Developments

The state-trace model approach, along with others like it takes many processes for granted. For example, one of the base assumptions made is that there is some consistent way of determining the content of an utterance and producing a semantic or propositional representation of this content; it is on this that the model works. I am assuming that at least the levels of abstraction shown in Figure 10.1 (taken from Winograd 1983: 17) are present in conversation and that the representation structures of any utterance can be determined. This is rather a sweeping assumption, but it is the basis of all natural language processing at any 'level' (i.e. in its simplest expression, that it is possible to process natural language systematically at all). The difficulty is in knowing how much interaction there is between these levels of abstraction in the interpretation of an utterance. How far can any part of speech understanding be truly abstracted from any other part?

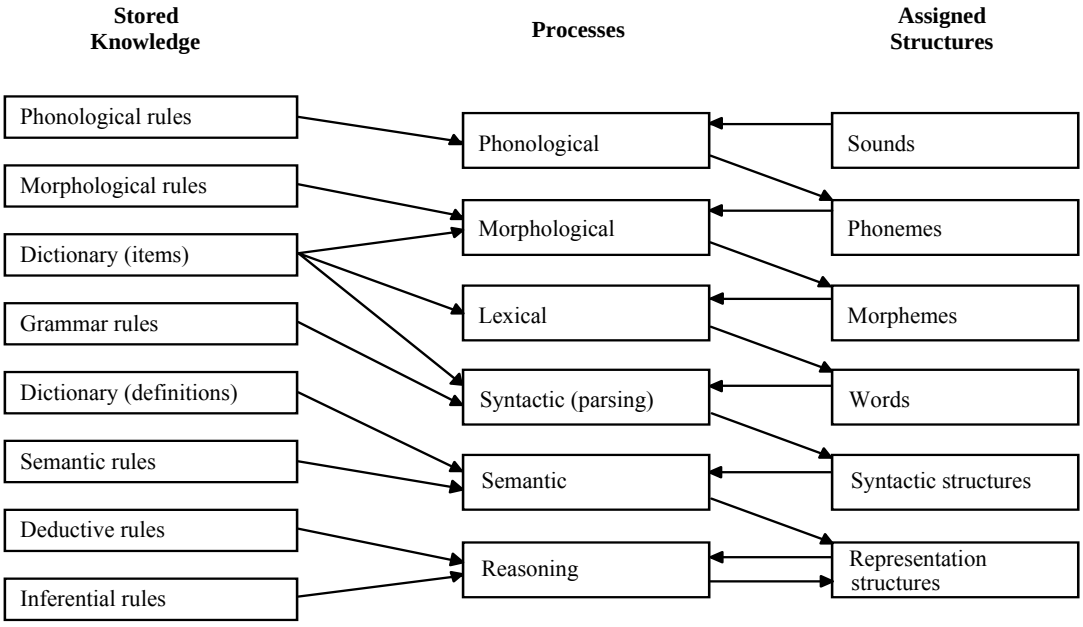


Figure 10.1 A stratified model of language comprehension

The answer is probably that at each level, hypotheses are made as to probable interpretations and the most likely is chosen before moving on to the next level of abstraction. If the analysis fails at any level, the utterance analyst backtracks to an earlier level of abstraction and tries another hypothesis. This continues until the best 'solution' (i.e. the most likely interpretation of an utterance) has been found. Thus the state-trace model is a vital link in this chain of inference, but is limited by the lack of interaction with other levels of abstraction. A fully integrated theory of natural language processing as I have just described it is still a long way off. How can we begin to tackle this limitation?

One of the major weaknesses of the model in its present state is the form in which utterance processing is undertaken. The whole model would be greatly improved by the addition of a parser and grammar, which could deal with elliptical utterances (such as anaphoric uses of 'yes' and 'no' for example) and isolate IFID (illocutionary force indicating devices) and other such linguistic markers. The inclusion of such a 'front-end' to the state-trace model would greatly increase its potential for speech act recognition, and would begin to take into account the phenomenon of drawing inferences at different levels of abstraction as discussed above. However, this is not a simple task, and would easily provide enough material for another thesis.

10.1.1 Proliferation of Speech Acts

One of the criticisms of speech act theory is the 'where does one stop?' factor. I have heard it argued that one can invent any number of speech acts, such as 'nag', 'fool', 'get at', etc. and that it would be impossible to characterise all of these in a conversation.

There are a number of replies to this particular criticism. The first is that there is something particularly odd about saying that 'nagging', 'fooling' or 'getting at' are kinds of speech act at all. If anything they are examples of different lexicon for a pre-existing speech act, but even this explanation does not really fit the feel of such 'acts'. You can tell that these behave differently when you try to use them explicitly:

* 'I nag you to paint the fence'

* 'I fool you (in)to going to the shops'

* 'I get at that the dog is ill' OR * 'I get at you about your failure to wash up'

They simply do not work well performatively at all. The way to deal with these types of act is to explain them in terms of the recognition of frames of behaviour. The performance of some speech acts under certain conditions (time, repetition, etc.) will trigger the identification of a particular behaviour. So, the framework for a nagging behaviour might be the repetition of the speech act REQUEST-ACCEDE more than once over a period of time. Now we can see that nagging is not a single speech act at all.

My position in this dissertation is that there are in fact relatively few classes of speech acts that one can perform, and that these speech acts are then recognised as instances of other kinds of acts or behaviour by the use of various IFIDs (Illocutionary Force Indicating Devices), mitigators describing one's attitude, etc., or by reproducing a particular kind of pattern, respectively. In this next section I will be arguing that even explicit speech act verbs should be treated as part of the lexicon to avoid such criticisms and so be treated mainly in the realm of semantics, not pragmatics.

10.1.2 Explicit and Implicit Illocutionary Acts

How is it that speech acts are identified in conversation? Often (but not always) it is the sentence mood or type that will indicate what kind of speech act is being performed (of course along with a variety of other different IFIDs, as discussed in Chapters 3 and 4). Table 10.1 shows an attempt to systematise the 'default' way of performing both explicit and implicit speech acts – if it is at all appropriate to talk about a 'default' given the caveats outlined and discussed in Chapter 4. In Table 10.1, V_I is the illocutionary verb (e.g. 'assert', 'command', 'promise', 'ask', etc.), and *c* is the content of the utterance.

Functional Type	**Speech Act Type**	**Explicit Format**	**Implicit Format**	**Response Type**
DESCRIPTIVE	*Assertive*	I V_I that *c*	Declarative sentence type. Often backed up with evidence of some kind, e.g. "Dad's coming home early tonight – I heard him tell Mum so".	*Assertive*
PRESCRIPTIVE	*Directive*	I V_I that you *c* I V_I you to *c* I V_I *c*	Imperative sentence type. Often performed indirectly for reasons of politeness, by being embedded in a question, e.g. "Could you pass the salt please?"	*Commissive*
	Commissive	I V_I (you) to *c* I V_I *c*	Declarative sentence type. The agent, or subject, is in the first person, the main verb is in future tense, e.g. "I will take the dog out for a walk".	*Directive*
(REQUESTIVE)	*Requestive*	I V_I you whether *c* I V_I you if *c*	Interrogative sentence type.	*Any*

Table 10.1 Explicit and implicit realisations of illocutionary acts.

It is interesting to note that one can almost always get to the implicit form of the speech act by stripping off the explicit illocutionary prefix. The process of identifying what kind of speech act the current utterance executes could then be shortcut by assuming that the act is explicitly named by the illocutionary verb. If all the preconditions for the named act are met within the current context, then this poses no problem. If they are not met, then a normal response might be to query the verb in question in some way, e.g. 'Surely you are not disagreeing, but agreeing?'.

This presupposes that the speech act is indeed explicitly designated by the illocutionary verb. If this were so, then a case when the utterance failed would be a genuine instance of incoherent conversation. However, Gazdar (1981) argues that the explicit illocutionary verb does not always indicate the speech act (just as the implicit illocution is not always fixed to the utterance). He gives the example of the uses of illocutionary verbs such as 'bet' and 'promise' (cited in Schegloff 1976: D11) for the purposes of attempting to close down a topic in a conversation. E.g.

'I promise you that'll be it'
'I bet that's it'

In these cases, the illocutionary verbs are performing the roles of strong and slightly weakened assertion respectively (see Section 10.1.3 for a full discussion of this phenomenon). The verb 'promise' here could be replaced by 'assure' with no significant loss of meaning, whilst 'bet' has the force of 'almost certainly' or 'most probably'. This use of 'promise' is interesting because normally it is used commissively with a world-to-word direction of fit, rather than assertively with a word-to-world direction of fit. Similarly, 'bet' is normally used as a declaration, with a double direction of fit (when the world is changed to fit the words in the instance of its performance). Note also that 'bet' is used figuratively or non-literally in the above example.

So both of these uses flout the 'default' interpretation of the verbs that they represent and cross 'types' of speech act. How are we to treat such examples? Do we say that they are not genuine instances of the explicit performative verbs? But clearly they are used performatively. It makes sense in these cases to say that some explicit illocutionary verbs have overloaded functions. Moreover, I would offer these kinds of examples as evidence that utterances using explicit illocutionary verbs can also be indirect and non-literal. For this reason, I would suggest that illocutionary verbs should be treated as part of the lexicon and be subject to the same defeasibility as implicit speech acts. If we treat illocutionary verbs as potential IFIDs, this would mean that we could finally eliminate the distinction between explicit and implicit illocutionary acts and produce a unified theory of speech acts covering all utterances.

10.1.3 Degree of Commitment

As discussed in the previous section, one of the first priorities for improving the model is to include syntactic and semantic processing that will allow utterances to be analysed in greater depth. The reason for doing so from the point of view of the development of the state-trace model would be the increased capacity of the model to recognise a wider range of speech acts.

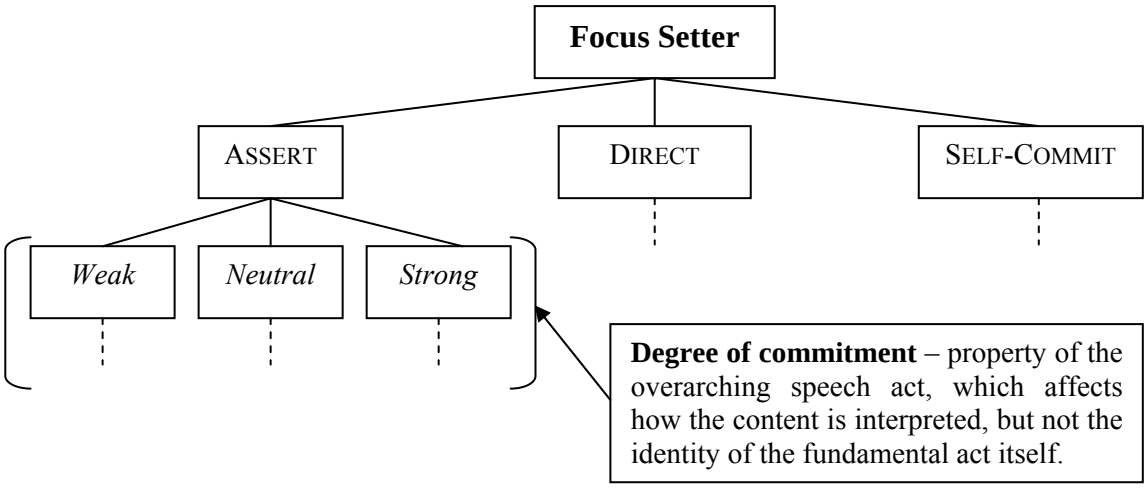


Figure 10.2 Diagram showing the speech act property 'degree of commitment'.

One means of doing this would be to include some representation of Searle's idea of *degree of strength* (see Section 3.2, and Sbisa 2001) – I would prefer the terminology **degree of commitment** as more appropriate for the state-trace model. The idea behind 'degree of commitment' is that a speech act such as *hypothesising* is actually a similar speech act to *asserting*, but with a weaker degree of commitment (see Figure 10.2). In other words, the basic **type** of the speech act does not change, but a speaker's **attitude**, or strength of commitment to the propositional content of the utterance does. Thus, if the commitment expressed by *hypothesising* turns out to be false, the speaker will be held less responsible for his commitment, because of the uncertainty expressed in the speech act. Similarly, the speech act of *swearing* that something is the case has a stronger degree of commitment than that of *asserting*; in this case, the speaker is held more responsible for his commitment than in the case of a straightforward assertion. Searle suggests that there is a scale of 'strength', roughly corresponding to the following:

$$\underset{-}{\text{HYPOTHESISE}} \quad \Leftarrow \quad \text{ASSERT} \quad \Rightarrow \quad \underset{+}{\text{SWEAR}}$$

However, this scale does not seem quite to capture the full complexity of degrees of commitment. In Figure 10.3 I have attempted to give a more finely grained scale of the level of commitment of a speaker.

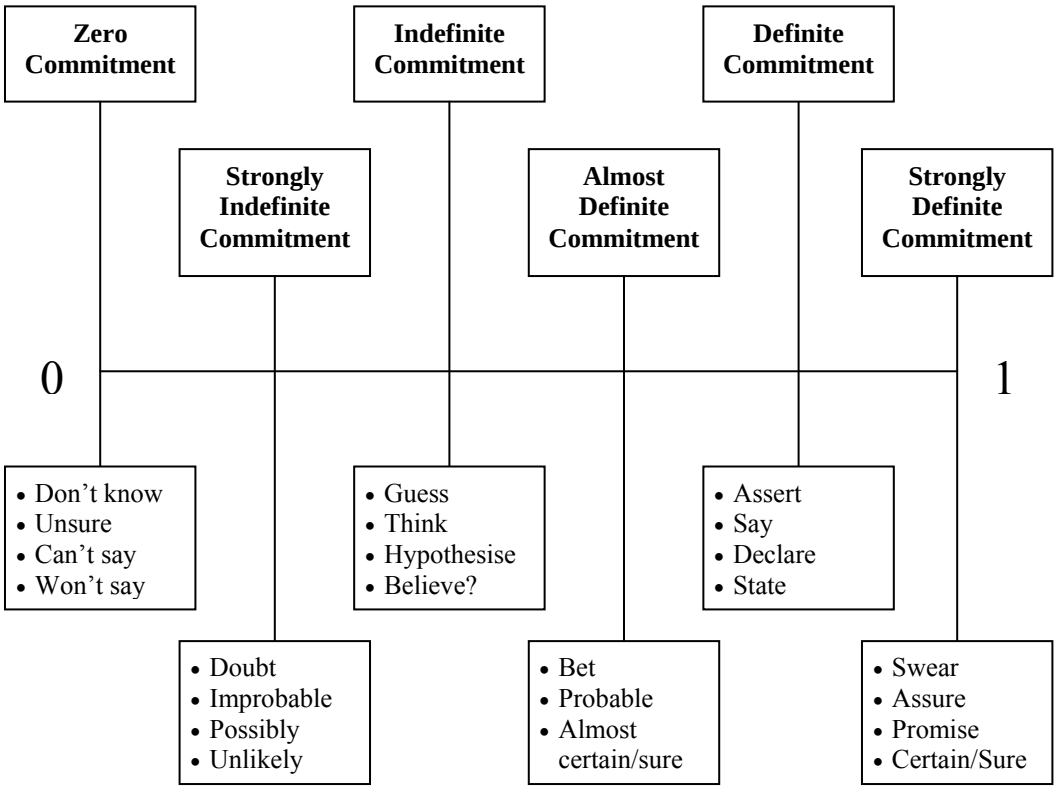


Figure 10.3 Scale of definiteness of commitment for ASSERT.

Sometimes assessing exactly where a speaker's attitude comes on this scale is rather hard. There are perhaps even finer shades of meaning than these, for example, 'most/highly probable' or 'extremely probable'. I have represented the scale in a linear way in Figure 10.3 for demonstrative purposes, but the degrees of commitment are in fact non-linear in nature. For instance, the list of speech acts that I have indicated as indefinite commitments are all probably more than 50% likely to be true in the mind of the speaker, otherwise they would weaken the definiteness further. Guessing that something is so implies that you have more reason to think that it is true than false. Similarly, if you state that something is so, you are expressing quite a high degree of definiteness; a hearer will have the right to assume that you have little doubt as to the truth of the proposition. But at the same time, it does not express as high a degree of definiteness as swearing that something is so. In a sense one wants to say that in both the cases of stating and swearing, the speaker is 100% committed to the truth of what they are saying. But perhaps in the latter case they are 101% committed? It is difficult to talk about percentages of certainty when dealing with language.

It is easy to slip into talking about degrees of certainty rather than of commitment, but this would be a misnomer. The degree of commitment expresses the public attitude of the speaker towards what they are saying; what they actually think, how certain they really are, could be quite different. In other words, it is not necessarily the speaker's real state of certainty. A speaker does not have to be lying to subtly alter what they are saying – the person lacking in

self-confidence may exaggerate their uncertainty, whilst the over-confident person (such as a typical salesman for instance) may exaggerate their certainty. The expression of someone's certainty has a lot to do with their character, with issues of social posturing and face, as well as with how important it is for that person to be believed (in the case of the salesman, his livelihood may depend on it), rather than with the true state of that person's 'sureness'. Further, there is even less correlation between a speaker's internal certainty and the actual truth of the proposition. It is for these reasons that I prefer to talk more accurately about the degree of commitment of a speaker to a proposition instead.

The vocabulary of the expressions themselves which denote degree of commitment are littered with allusions to mathematical probability: 'probable'/'improbable', 'likely'/'unlikely', 'certain'/'uncertain', 'possible'/'impossible', 'definitely', 'bet', 'it's a safe bet that', etc. There is no doubt that there is a strong relationship between a person's level of commitment to a proposition (or action) and that person's perception of the likelihood of its truth. However, it can plainly be seen that certainty and commitment are incommensurate when we look again at the scale as depicted in Figure 10.3. Perhaps somewhat counter-intuitively I have defined zero commitment to be the expression of the speaker's complete uncertainty or inability to say whether the proposition is true or false; zero commitment is the refusal to commit to either position, not the zero probability of the truth of the proposition.

The reason for this is that, as discussed in Section 9.2.1, I take the proposition to be just as likely to be in the negative as it is to be in the positive – I treat X and ¬X in exactly the same way. So in my model, varying the degree of commitment would produce the following negative speech acts (here X is a proposition such as 'Manchester United is great'):

I swear that ¬X		
I assert that ¬X		
I think it's probable that ¬X	≈	I think it's improbable that X
I guess that ¬X		
I think it's improbable that ¬X	≈	I think it's probable that X
I don't know if/whether ¬X	≈	I don't know if/whether X

Note the unsurprising equivalence of some of the options (showing the flexible scope of 'not' with respect to commitment mitigators). Some of these are more often performed in one form rather than another in order to avoid the double negative.

Some speech acts are inherently negative, for example: 'deny' ≈ 'assert not', 'assert' ≈ 'deny not' and 'forbid' ≈ 'allow not', 'allow' ≈ 'forbid not'. In the model as it stands I have treated these as if they behave in a similar way to other speech acts, just with the opposite negativity. This seems to work well enough: 'I wasn't at his house on the night of the murder' might

equally be reported as ‘She denied that she was at his house…’ or ‘She asserted that she was not at his house…’. However, I am not entirely convinced that the original utterance could just as well be rephrased ‘I deny that I was at his house…’ as ‘I assert that I was not at his house’; the former implies a certain defiance that is perhaps absent in the latter – although ‘I tell you that…’ also expresses defiance, but with a slightly different nuance (of insistence). There is obviously more to be discovered about the use of negative speech acts – the phenomenon would provide an interesting topic of research in itself. I digress.

Although I have been discussing the various degrees of commitment for the speech act ASSERT, this idea works just as well with other speech acts. Consider the difference between the following:

‘Well, perhaps you’re right’	ADMIT
‘No, you’re right’	CONCEDE
‘No, you’re absolutely right’	COMPLETELY CONCEDE

The addition of an indicator of degree of commitment to the speech act schemata will increase dramatically the number of different speech acts that can be identified. Utterances containing illocutionary verbs will be assumed to have the degree of commitment to the content explicitly expressed by the illocutionary verb (subject to the kind of defeasibility discussed in Section 10.1.2). All other utterances will have to be judged on some other criterion; the most likely indicators of degree of commitment in implicit speech acts, are linguistic markers such as ‘definitely’, ‘presumably’, ‘possibly’, ‘I am not sure’, ‘I am sure’, ‘I am certain’, ‘probably’, ‘maybe’, etc. These all modify the *zero* degree of commitment in some fairly definitive way. The grammar would have to be able to parse such lexical markers before the concept of ‘degree of commitment’ could be properly covered. Adding the feature of degree of commitment might substantially increase the complexity of the preconditions of each speech act schema; extensive work would have to be carried out in order to redefine the schemata and to check that all the relevant speech acts were still correctly identified. Developing this dimension of the state-trace model would help distinguish between a range of speech acts hitherto largely ignored by most researchers working in disciplines in some way connected to speech act theory.

10.1.4 Beneficiary Factor

Some kinds of speech act cannot be identified at all without recourse to semantic knowledge. However, I hope to show that even these acts are modifications of the underlying basic speech act, and that rather than being speech acts in the sense that I have defined them, they are actually behaviours that are recognised and performed by using speech acts in a particular manner. Many more types of ‘speech act’ could be characterised by the ability to recognise in whose benefit or to whose detriment the fulfilment of a speech act would be (Stenström 1994). This

also ties in with some of Leech's (1983) observations about politeness and direction of benefit, as discussed in Section 4.3.

The speech act interpretation of a SELF-COMMIT commissive utterance is altered by the direction of benefit or detriment of the proposition. For example,

'I'll wash the car for you'

is interpreted as an OFFER because the addressee will enjoy the benefit of the completed action. However,

'I'll punch your nose for you'

is interpreted as a THREAT because there is no benefit, and potentially quite a severe cost, to the addressee on completion of the action. Note though that they are both still acts of self-committal.

The speech act interpretation of a DIRECT directive utterance is altered in a similar way. Take the very Yorkshire example,

'Put ' kettle on'

which will be interpreted as an ORDER because it is the speaker not the addressee who will benefit (even though both parties might eventually benefit from the making of a cup of tea). But,

'Go and have a lie down'

is interpreted as a SUGGEST act because it is in the addressee's interests to comply.

	COMMISSIVE	**DIRECTIVE**
speaker BENEFIT	SELF-COMMIT *I'll shut the door*	DIRECT/COMMAND/ORDER *Make me a cuppa*
addressee BENEFIT	OFFER/PROMISE *I'll carry that for you*	SUGGEST/ADVISE *Go home*
speaker DETRIMENT	? *I'll slit my wrists*	SACRIFICE? *Have the last cake*
addressee DETRIMENT	WARN/THREAT *I'll make you eat it*	? *Go jump off a bridge*

Table 10.2 The effect of 'benefit' and 'detriment' on speech act interpretation.

Some of the different options for benefit and detriment are plotted out in Table 10.2. It is almost certainly not the exhaustive list, as there are further combinations that I have not accounted for (the difference for example between an act where both the speaker and the hearer

benefit and where the speaker benefits to the addressee's detriment). What happens when neither party benefits, or where it is unclear who benefits? This is not supposed to be a rigorous list, but show that various different speech acts can be identified by the inclusion of this feature (see also Tsui 1994: 104-108 for a discussion of how direction of benefit can affect the identification of different requestive types of utterance).

Some of the speech acts shown in Table 10.2, such as SACRIFICE, would never be used as illocutionary verbs in conversation. If I said, 'I sacrifice my last cake to you', this would be taken as rather a funny thing to say – somewhat jokey, assumed melodrama. However, we do recognise the behaviour represented by SACRIFICE because of the direction of benefit – i.e. benefit for the addressee, but detriment to the speaker (they get no cake).

Many utterances are ambiguous in terms of benefit and detriment; interpretation is very often extremely context and situation dependent. The utterance, 'Go to bed' will have the force of SUGGEST if spoken to a tired grown-up, but of an ORDER if addressed to a naughty child being punished (although arguably it might be to the addressee's benefit). Actually, these interpretations might have more to do with the authority of the speaker than anything else; but even so, we do interpret what someone says differently depending on our perception of what is being asked of us. Sometimes this very ambiguity can be exploited in order to fool someone into doing something that's more in one party's interests than another.

To recognise the difference between a PROMISE and a THREAT requires knowledge of the world, but the two are demonstrably closely related. They are of the same functional type. This in part explains the non-literal uses of their illocutionary verbs:

'I promise you that I'll get even with you' = THREAT
'I warn you that I'm going to hug you' = PROMISE?

This occurs when the content does not match up to the expected direction of benefit suggested by the verb.

Many of these kinds of speech act are hard-coded in conventional turns of phrase which signal their performance:

'I would suggest'	SUGGEST
'Why don't you…'	
'You could…'	
'You might want to consider…'	
'If I were you, I would…'	ADVISE
'If you want my advice…'	
'What I'd do is…'	
'If I were in your shoes I would…'	

‘Shall I…?’	OFFER
‘Would you like me to…’	
‘…if you like’	
‘…if you’re not careful’	THREAT

Combining these kinds of speech act with what I call degree of commitment will produce other speech acts, e.g. an OFFER with a strongly definite commitment would become a PROMISE, and similarly SUGGEST has a weaker degree of commitment than ADVISE. In these cases, the term ‘degree of commitment’ does not fit quite as well as ‘degree of strength’. Perhaps for prescriptive utterances it makes more sense to talk about a difference of strength of attitude towards the prescription? Or perhaps these are two separate dimensions? For instance, when a COMMAND is issued as a REQUEST, is there any way that a speaker is less committed to the prescription? The COMMAND is weakened only in the sense that there is a tacit recognition that the addressee in complying would be doing the speaker a favour.

10.1.5 Partial Speech Acts

There is a further dimension to the interaction of speech acts for which I have not yet accounted in the model presented here, namely the performance of partial speech acts. These are acts that agree or disagree (arguably both at the same time!) with a part of a proposition, but challenge another. They mainly occur in complex propositions; one cannot partially agree with ‘I am a girl’, or at least not seriously. However, there are some facts that are subject to change or differing states. For example, ‘The sky is blue’ is true sometimes and not others. If spoken as a generalisation, one could partially agree with it by saying, ‘Yes, but not always’. There are a variety of ways that partial speech acts can be performed.

If I assert that $(P \wedge Q)$ is true, we can treat this assertion in at least two different ways. We might separate the two into their two sub-parts ‘I assert that P’ and ‘I assert that Q’, so that when the next speaker says ‘I agree that Q, but not P’ we also treat these separately and can say that the second speaker is agreeing with one proposition and disagreeing with the other. Alternatively, if $(P \wedge Q)$ is treated as one entity then what the second speaker says must be interpreted differently. Logically, if $\neg P$ is true, then the term $(P \wedge Q)$ is not true. But there is more going on in conversation than this. The second speaker is not merely suggesting a replacement for $(P \wedge Q)$, i.e. $\neg(P \wedge Q)$ because $\neg P$, but a new term, Q. So as can be seen, there is a sense that speakers can partially agree or disagree. Anecdotal evidence for this would be the fairly common use of the phrases ‘Yes but…’ or ‘I agree with you in part/partially, but…’. Let us clarify the problem with an example:

A: Cath and Jim are supposed to be coming out with us tonight.

B: Well, Jim probably will, but Cath's working. She's just told me so on the phone.

A: Oh, that's a shame – I haven't seen her for a while.

So, we can see that while the proposition as a whole (that both Cath and Jim will come out tonight) is claimed as untrue, **B**'s reply cannot be simply seen as a disagreement because they are also claiming that part of it is true; it is also a partial agreement.

I have been discussing partial agreements here, but there are also equally partial concessions, partial acceptances, partial accessions, etc. In fact, all referring speech acts can be performed partially on complex propositions.

In many ways, it could be argued that the first approach is theoretically more appealing. This is because often the performance of a partial act can provoke further reaction and discussion concerning the sub-proposition being disagreed with. However, the resolution of the sub-proposition will affect the resolution of the whole.

The performance of some partial speech acts relies more heavily on the division and dependencies of the sub-sections of the original proposition. For example, there are speech acts that deny the premises while accepting the conclusion:

A: Cath's got too much work to come out tonight.

B: She hasn't!

This implies that either there is some other reason why Cath is unable to come out, or that there is no reason why she could not come out. Here **A** is saying $(P \rightarrow \neg Q)$ – where P is the proposition that Cath has too much work and $\neg Q$ is the proposition that Cath will not come out tonight – and **B** is claiming $\neg P$, which leaves $\neg Q$ in question. Conversely one can also deny the conclusion while accepting the premises:

A: Cath's got too much work to come out tonight.

B: Yes, she does have too much work, but she's coming out anyway.

Here **B** is not denying the truth of P, but saying $\neg Q$ regardless.

The interpretation of the above examples could be accommodated in the model of speech acts put forward in this dissertation by the development of logical rules to govern the manipulation of propositions. Addressing this issue could well open a can of worms though as there are all sorts of logical combinations that are largely left unaccounted for in current speech act theory as a whole.

I have tried to show that partial speech acts could be absorbed by and are not incompatible with the model as it stands, but this is too large an area of inquiry to cover comfortably here and falls beyond the scope of my current research. A future expansion of my model might look at this phenomenon and incorporate a propositional logic for speech acts. Further work might also try to incorporate and integrate the following: epistemic logic (relating to knowledge and to its degree of validation), modal logic (relating to the mood of the main verb), temporal logic (relating to how the proposition is anchored in time and the tense of the main verb will affect the identification of its function[1]), deontic logic (relating to duty) and doxastic logic (of or pertaining to opinion). At present, the model brushes over the difference between a refusal that is phrased "I will not do X", "I should not do X" and "I cannot do X".

10.1.6 Some Problems with General Conversation

In Chapter 7, I outlined the reasons why I decided to focus on conversational discourse in order to search for a generic model of speech acts. In retrospect, there were a number of serious drawbacks as a result of this choice.

Looking at general conversation was a bit ambitious. It is not the 'cleanest' of dialogue types in terms of turn-taking, formality of grammatical constructs, etc. There are also many subtleties such as social niceties, extra layers of meaning (due to the participants' knowledge of each other), shortcut inferences, jokes, extended narratives, and so on, all of which make functional analysis difficult and challenging. Choosing general conversation, where participants are often chatting for no more complex purpose than to establish a social relationship with another human being by relating and sharing information, was perhaps not the best choice. The kinds of speech act used were limited. Finding extensive examples in order to prove the hypothesis that there is a basic, simple interactive structure that underlies all dialogue was somewhat challenging.

10.1.6.1 Disproportionate Number of Assertions

Although it seemed like a reasonable idea to look at the kind of interaction type that is the fundamental building block of all other dialogue situations, in reality the interaction that takes place is largely of one type – it is assertive in nature. The preponderance of these speech acts in conversational discourse was also found in the SWBD-DAMSL speech act annotated corpus. In the analysis carried out by Stolke et al. (2000) – see Table 6.2 – 36% of all the speech acts annotated are of the type STATEMENT (equivalent to ASSERT in my model). I have replicated

---

[1] See Shanahan (1999) for a description of event calculus.

similar spreads of percentages[2] in the collection of my own corpus (although they are not directly comparable due to using different paradigms). Certainly the number of ASSERT speech acts far outweigh the occurrence of other speech acts in my sample data. Perhaps this was an accident of the kind of conversations I recorded, although it does seem to be a feature of conversation in general.

In future it might be of more interest to study a variety of different types of interaction to draw comparisons between and conclusions about the distribution of speech acts in these different types of dialogue. This is not an easy thing to do however considering the kinds of restrictions that I suggest should be imposed upon the collection of such data, and the overall difficulty of collecting it anyway (see Chapter 7 for a comprehensive discussion of these problems).

One idea for some future investigation would be the application and testing of this interactive model on child-adult conversations. If, as is my hypothesis in this dissertation, people acquire this basic interactive model by years of learning language through conversations with their parents, siblings, teachers, etc., then there should be evidence of a greater variety of speech act types in the dialogue of young children. This is backed up by children's tendency to practice transactional situations (such as going shopping) through their play. However, this idea is mere speculation and would require extensive investigation.

10.1.6.2 Justifications and Explanations

Often in spoken dialogue people will support any claim or request that they make (Tsui 1995). This is especially true (but not exclusively so) when they are responding negatively to someone else's claim or request. For instance, if we revisit the example conversation in Conversation 1.1, Chapter 1, we can see that when Albert refuses June's order to shut the door, he does so by saying 'I can't' not 'I won't'. This is not just a simple refusal but also implies that he has a good reason why he is unable to comply. There is normally an assumption that the hearer will follow the speaker and adopt their goals if appropriate, or not too inconvenient, or not contrary to their own goals. How should explanations and justifications for an attitude taken up by one of the participants in a conversation be represented in my model? This is a difficult question to answer. In some senses it is a very special case of repetition, in that, in general, it provides a logical implication of the 'content' of a previous speech act. E.g.

---

[2] Note that I do not have exact figures to present here, mainly because I do not have enough data to extract information of statistical significance.

A:	Shall I clear this up?	*Clear_Up(A, this)?*
B:	No,	*¬Clear_Up(A, this)*
	I haven't finished yet thanks.	*¬Finished_With(B, this)* *[→ ¬Clear_Up(A, x)]*

This is a product of a deep reasoning process. In order even to know that **B**'s not having finished with whatever 'this' is causing a mess, one has also to know that if **A** clears it up, **B** will be unable to finish what he is doing. To understand that 'I haven't finished yet' is a justification of a rejection of an offer requires a level of semantic and world reasoning that is far from simple or straightforward. I would suggest that this is not the business of the high-level, structural model that is under discussion here, although it should be clear that trying to account for such a phenomenon would not produce any inconsistencies. One could say that 'I haven't finished yet' is the assertion of a new fact (which it clearly is also), but because of its position in the conversation coming just after declining an offer of help, and because there is usually an expectation of some sort of explanation or commentary to punctuate and clarify why we perform the speech acts that we do, it is interpreted as a justification.

Justifications happen quite commonly in almost any kind of dialogue and not just when breaking bad news to one of the participants. Extra information to support what one has said can be given after any speech act, positive or negative.

Positive:

Mandy: Oh, switch it off. I'm going to have to do it on unsuspecting… [people]
①

Negative:

Linda: I need details, I need details

Mandy: No, no, I can't, I can't,

Linda: Yes you can

Mandy: Not, not while I'm recording ②

Linda: If anyone's confidential about it you will be

Often explanations and justifications are preceded and cued by the word 'because' to explicitly signal the purpose of the speech act to the hearer. Arguably 'because' is implied in many sentence constructions and the utterance could be rephrased to include it. If we look at the two examples above:

① ⇒ 'Switch it off because I'm going to have to do it on unsuspecting people.'

② ⇒ 'I can't because I'm recording'

Explanations can also be embedded within the original assertion to pre-empt a potential objection:

> **Mandy**: But, I think, in order to get unstilted natural conversation, um, there's a case to be made for recording people without their knowledge.

This puts paid to the idea that explanations are a form of repetition. They seem to be an integral part of the case for the speaker's position.

Explanations and justifications are analysed as new assertions in my model as it stands, which, although not incorrect, is obviously not ideal. With the addition of some kind of semantic processing and logic, the fact that this assertion is also an explanation for a prior speech act might be inferred. So, they are not really either assertions or repetitions but support for a previous act. Sadly this interpretation is beyond the scope of the model at the moment.

10.1.6.3 Stories

Closely related to the problem of what to do about explanations, is the problem of how to account for the propensity towards storytelling in conversation. Sometimes stories are told as a kind of justification or illustration of a previous claim or action, in which case the sequence of events will be something like the following (which is very similar to the structure of explanations given above):

Claim (X)
Story (Y) [→ Claim (X)]
[∴ Claim (X)]

An example of this is shown below from one of the conversations that I collected; here the discussion is about whether it is possible to record people's conversation using an answer-phone.

> **Mandy**: no, yeah, but you know the actual machines you plug in.
>
> **Nigel**: yeah, because often like my mum's one it's so stu-, I hate it 'cause it sort of (pause) if you don't get to the phone in time it comes on -
>
> **Mandy**: yeah.
>
> **Nigel**: while you're, like my mum on the pho-, I'll be trying to talk to my mum, and then the the answer phone comes on. and mum's li- "oh I've got to switch it off", so yeah it would be recording our conversation, yeah.

However, sometimes there seems to be very little actual point in telling a story; the point is the story itself. The most that one can say is that the storyteller is in some way committing himself to the truth of the story he is elaborating.

As in the case of justifications, stories provide a big problem, not just to the working of my model, but also to speech act theory itself. What is it that we are doing in the act of storytelling? Is it simple information passing so that close members of a social group will be possessed of similar (though of course not identical) information? Is storytelling a form of entertainment, a means of conveying surprising or unexpected information? Is storytelling used in order to gain a closer relationship with acquaintances who are not quite so well known to us, to align and position ourselves socially? Or is it a combination of all of these things?

When people tell stories, there is almost a switch of mode (Clark 1996). The stories are encapsulated entities within the conversation and the focus is entirely subsumed by the 'point' of the story. In fact, even stories that are not told in support of some claim are expected to be relevant in some way to the conversation or to the other participants. People are constantly asking themselves while they are listening to the story, 'Why am I being told this? What is the point? What conclusions should I draw from the information provided?', even if doing this is merely in preparation of an appropriate response to what they are being told.

I will not delve any further into stories and storytelling here, where they are only a side issue with respect to the central thesis. For more detailed discussion of this phenomenon, see Jefferson (1978), Tannen (1984) and Laurenceau et al. (1998). I am not proposing to deal with storytelling within my model as this is beyond its current scope; however this particular feature of conversation is very interesting and deserves further investigation.

10.1.7 Other Potential Research Directions

In this section I shall briefly summarise some of the other directions that the STM could be taken in order to increase the range of speech functions characterised by the basic model. Some of these extensions have been discussed at length already.

10.1.7.1 Beliefs, Plans and Goals

Having described the context for identifying a wide range of speech acts, a suitable next stage would be to attempt to incorporate some modelling of beliefs and reasoning. At present the STM only deals with speech act *recognition*, and while there is still a long way to go with this function of the model, I see it eventually incorporating speech act *generation* as well. This may well include looking at plans and goals in conversation; there is a substantial body of work in this area already, so it should be possible to adapt the theories in existence to cope with the STM architecture. A final aim would be to have several autonomous agents endowed with differing belief systems conversing with each other according to their STMs and beliefs, backtracking when the conversation shows model inconsistencies, updating commitments constantly, but yet allowing each agent/participant's model to 'come apart' as it were from each

other's (accounting for misunderstandings and contradictory belief systems in 'real life'). This idea is very much in line with current Agent Negotiation Protocol and Theory in Distributive Artificial Intelligence systems, but I suspect quite a long way off from realisation in any true to life model of conversational agents.

10.1.7.2 Literal and Non-Literal Speech Acts and Humour

Very much related to 10.1.7.1 would be work to understand the force of non-literal speech acts, both in general conversation and in interpreting humour. According to Bach and Harnish's (1979) schema to deal with literal, non-literal, direct and indirect speech acts, we should be able to account for the following:

Literal direct:	'Close the door' *Direct command, surface imperative*
Literal indirect:	'Would you close the door?' *Indirect command, surface interrogative*
Non-literal direct:	'Open the door a bit wider, I'm not cold enough' *Direct command, not literal*
Non-literal indirect:	'Were you born in a barn?' *Indirect command(?),not literal*

I have argued that there are no indirect speech acts, and that these are explained by direction of fit. So in order to interpret non-literal speech acts, we must be able to reason about the content of an utterance.

10.1.7.3 Error Correction in Natural Language

I am very interested in the way people notice mistakes in a conversation and how they try to recover from them (Jefferson 1974, McRoy and Hirst 1995). Much of the time it is the failure to provide the 'right' answer that alerts participants that some kind of error has taken place. This extends to speech acts too (you can misinterpret the intended speech act and respond inappropriately for that reason):

(1) **Mandy**: Steve, are you going home any time now?
(2) **Steve**: I'm very happy for you – I'm going for a curry with Sarah.
(3) **Mandy**: No, I said: "Are you going home?".

Steve has misheard (1) as "Steve, I'm going home any time now", and interprets this as an assertion instead of as the question intended. The ability to identify and recover from such errors provides very clear indications that there is an expectation of co-operative, structured behaviour in conversation.

10.1.7.4 A Functional Record of Conversation

People are very bad at remembering things verbatim. Generally they only take away a functional representation of a conversation (as I argued in Chapter 8). The STM reflects this in that it updates the current state of the set of spoken commitments in a conversation and keeps a chain of reference to other speech acts performed in the same context space. People reconstruct conversations all the time, especially when a new participant joins the conversation and wants to know what the current participants have been talking about. It would be an interesting project to see how this might be achieved using the model.

10.1.7.5 Modelling Trust and Distrust

The STM could be used for the cognitive modelling of confidence in a person (Castelfranchi et al. 1999, Bickmore and Cassell 2001). A person's trust or confidence in a speaker would go up or down depending upon whether the speaker has honoured his past commitments or not. Clearly modelling trust would affect the interpretation of a person's utterance. For example, if I say 'I will meet you at the pub at 8pm' but have been known not to honour my commitments in the past, this utterance might be interpreted as 'I will probably meet you at the pub at 8pm, but more likely it will be sometime between 8.30-9pm'. This is an idea that is obviously beyond the scope of this dissertation, but nevertheless would follow logically from the model of cognitive processing that I am proposing.

This idea of modelling trust is also closely related to modelling beliefs, and degree of commitment. What we trust and do not trust about what people tell us has an impact on the acquisition of our beliefs. How does hearsay change from a piece of information for which we lack evidence to what we would consider knowledge? Is it the repetitive affirmation of the proposition over time? Do we accept what we hear until we have evidence to the contrary? Is it the source of the information that sways our degree of certainty?:

'I heard somewhere that…'
'Bill told me that…'
'I read in the paper that…'

These could all be methods for absolving oneself of responsibility for the accuracy of the information and of blame if it is wrong, and therefore weakening one's commitment; or another point of view is that they are ways of bolstering up the proof of what is said and are therefore strengthening one's commitment.

10.1.7.6 Incorporating Gestures and Multimodal Communication

Technology has moved on from mere recordings of sound; now moving images can be recorded and analysed as well. It is now possible not only listen to spoken language as it is uttered, but also to examine the visual and contextual clues that aid in the interpretation of communication. A whole new discipline investigating multi-modal interaction has started to build up around this and other new technologies (Serenari et al. 2002). Interactive and mixed-initiative three dimensional computer graphics modelling (such as Cassell et al. 1999, Cassell 2000) has led to the development of computational spoken dialogue agents which use artificial intelligence vision techniques to recognise gesture and facial expressions, as well as using direction of gaze to help to indicate when and who should take the next turn. These visual signals augment the ability of the virtual person to respond relevantly to spoken instructions.

10.1.7.7 Language Independence

One interesting idea for future research is to find out whether the basic structure of conversational discourse that has been developed during the course of this project is generic to all languages. My hypothesis would be that it is, but that the realisations of the more specific speech functions are not.

It can be seen from the discussion above that there is plenty of room for further developments to the state-trace model, the strengths and weaknesses of which have been brought out and highlighted by the current work.

10.2 Achievements

In this dissertation I have tried to show that the theory of speech acts has much to offer to those who seek a means of interpreting utterances in a conversation (or indeed any other spoken interaction between human beings). My approach has been based on the premise that, 'Human beings can fairly reliably recognise speech acts and therefore there must be some mechanism by which it can be achieved'. I do not pretend that I have correctly and sufficiently isolated such a mechanism, but I think that I have given some idea of its 'shape' or 'form', and of how this might later aid language understanding from a computational point of view.

Having covered at some length how the model could be expanded to account for some of its limitations, I will now briefly outline what I believe are the main achievements of the project before drawing the discussion to a close.

10.2.1 Aims and Objectives of the Thesis

By and large, the original aims set out at the beginning of this dissertation in Section 1.4 have been fulfilled; where this is not the case, I have indicated so in the previous chapters, and given the reasons why not. The objectives that I identified were:

(1) To account for the fact that sequences of speech acts have dependencies of more than one conversational turn. I have done this by developing a model of referential speech acts that rely on the identification of previous speech acts and the state of the prevailing context space for their 'correct' analysis.

(2) To highlight the difference between the underlying act, or function of the utterance, and the behaviour that it represents in context.

(3) To show that the theoretical model can be applied to and validated against real conversational data.

(4) To develop a grammar of discourse, which could later be semantically and pragmatically 'clothed' or embodied in spoken dialogue (see Section 10.3).

(5) To show how conversation has to be modelled from one, single, yet role-changing, perspective, because language understanding is essentially solipsistic in nature. This allows for different participant's models to come apart. One potential method for dealing with this phenomenon would be to 'take back' or 'rewind' context spaces to a prior state in order to find out where an interpretation has gone wrong. Although I have not explicitly outlined the mechanisms by which this might be done with my model, I think it is clear that by undoing the moves that updated the context spaces, we can return to an earlier state of play and reanalyse the subsequent moves (cf. chess analogy in Chapter 8).

(6) To show that a speaker's underlying intention should not be a consideration for the felicitous performance of a speech act – the only thing that counts is how the act is interpreted in the context and to what the speaker is committed. A speaker might not intend to perform an action to which he commits himself, but the hearer has no way of knowing that. This shortcuts the need to start reasoning about what another person believes and intends us to believe when analysing a conversation.

The state-trace model characterises many distinct types of speech acts from the utterance and the context combined. These are roughly representative of the basic speech act types, which can later be used in the characterisation of other speech acts. I have shown that there is a method by which the function of utterances in a conversation may be processed consistently and systematically to identify the speech act it performs within a conversational context.

10.2.2 Scope of the Thesis

It is interesting to note that in the preamble to her book about different approaches to discourse, Deborah Schiffrin (1994) asks the question, "How can we define discourse analysis in a way that captures it as a field of linguistics and differentiates it from other studies?". She then continues to explain how she will answer this question: "I suggest that two prominent definitions of discourse (as a unit of language larger than a sentence, as language use) are couched within two different paradigms of linguistics (formalist, functionalist)".

To me this underlines two points of view that pose serious obstacles to coming to a total understanding about how we use language to communicate. The first is a kind of speciesism: because discourse is language, it must have its explanation rooted in the field of linguistics. I would argue that you cannot divorce discourse from other theories such as cognition, mental models, agency, belief structures, action and socio-ethnological factors too. This makes it very difficult to keep all the issues in play at the same time, but it is a mistake to think that any approach could be complete without at least addressing how it fits in with other related areas.

The second is a kind of paradigm favouritism: to have the correct approach you must either be in the formalist or functionalist camps. Again I think this is a mistake (and unfortunately a common one – see for example the scathing rejection of formalist approaches in the preface to Wilks 1999). There is an undeniable dependency between form and function. Again theories that fail to incorporate or attempt to explain how they are positioned within both camps will not succeed in generality of application.

	VERBMOBIL	STM	SWBD-DAMSL
(1) Specific/Generic?	Specific	Generic	Generic
(2) Open/Domain Restricted?	Domain	Open	Open
(3) Task-Driven/General?	Task	General	General
(4) Observed/Structured?	Structured	Structured (and semi-validated)	Observed
(5) Speaker-Hearer Distinction?	×	✓	×
(6) Acoustically Informed?	✓	✓/×	×
(7) > 2 Speakers?	×	✓	×
(8) Discontinuous Data Dependencies?	×	✓	×
(9) Levels of Abstraction?	×	✓	✓/×
(10) Semantic/Pragmatic Distinction?	×	✓	×
(11) Syntactic/Pragmatic Distinction?	✓	✓	✓/×

Table 10.3 Comparing the theoretical distinctions of VERBMOBIL and SWBD-DAMSL dialogue act annotation schemes with my speech act model.

I have attempted in my work to avoid, as much as it is possible to do so, the two pitfalls of speciesism and favouritism. In Chapter 1, Section 1.5, I outlined a list of desiderata for the study of speech acts in conversational discourse. At the end of Chapter 7, Section 7.3.3, Table

7.5, I measured two of the leading dialogue annotation schemes (which represent the two extremes of generality and specificity) against this list. In Table 10.3, I show how the theoretical model that I have developed during this research fits in with the VERBMOBIL and SWBD-DAMSL schemes. I have tried to address the inadequacies that were noted with these and fulfil the specifications laid out in Section 1.5.

Overall, although the STM would require a significant amount of work still before a computational system could be built based on it, in my opinion it bridges the gap between a variety of different theoretical positions and formalises some of the structural observations made in such subjects as conversation and discourse analysis.

10.3 End Word

What I have developed here is the skeleton model of conversation, the bare bones of a theory of speech acts. As such, it will seem very simplistic and almost 'undressed'. It is hardly surprising that the examples will seem as though a lot of information is lost by such a simple description as provided by my schema of interpretation. My model would not be an adequate representation if I intended that it be used as a standalone. But I do not. It was my intention to strip the problem down to its basic constituents in order to be sure that none of the foundational functional structure was missing. I wanted to be sure that the speech acts that depend on the performance of other speech acts were clearly defined; that the search tree for speech acts was complete as far as possible or reasonable, that every node was traversed. This I have done with the aim of then showing how this skeleton might be embodied and clothed to appear more like natural conversation, but without losing any of the underlying principles and structures that give the functional meanings the shape on which to hang more finely tuned interpretations.

My aim was to start to map out a grammar of discourse, to explain how it is that we understand what we mean from what we say. This aim I feel I have achieved. Not that I believe that this is the whole story: just as grammarians looking for a definitive sentential grammar have not yet found the set of rules that comprehensively accounts for every sentence formulation available in any language, no-one working in dialogue modelling has yet managed the same at the utterance level. However, many grammars are sophisticated enough these days to be able to encompass the overwhelming majority of (grammatically correct) sentences, to the satisfaction of a given set of criteria. One of the factors influencing the effectiveness of sentence grammars is that the complexity of the problem of coverage of different formulations increases exponentially depending upon the fineness of 'grain' of the grammatical categories. The same can be said of discourse grammars. I have tried to coarsen the 'grain' in order to make the problem of identifying legal sequences of speech acts more tractable. It remains to be seen whether this abstracts away so far from the original conversation that the resultant analysis is worthless, or

whether it provides a useful framework and basis from which to develop a more intricate and comprehensive discourse grammar.

Work on the development of the model presented in this dissertation has been an exciting, interesting and challenging project. Perhaps in retrospect the aim of the project to plot out a discourse grammar based on conversational speech was a little over-ambitious. There were so many different directions in which the research could have conceivably gone that arbitrary decisions had to be made at times. The scarcity of exactly similar work in this particular area of research led to difficult decisions concerning the representation of the model. On the other hand, the bewildering and prolific range of work in closely related fields made a thorough literature review difficult to carry out. Conscious of this, I have spent a lot of time justifying the choices I have made at each point, and also in adopting and adapting methods and principles from related fields whenever possible. I believe that this preliminary exploration of the applicability of the model to example conversation has yielded some promising results and points to further interesting work to expand the model and to address the various limitations I have outlined here.

Bibliography

AITCHISON, J. 1989. *The articulate mammal: An introduction to psycholinguistics.* London: Unwin Hyman.

ALEXANDERSSON, J., BUSCHBECK-WOLF, B., FUJINAMI, T., KIPP, M, KOCH, S., MAIER, E., REITHINGER, N. SCHMITZ, B. & SIEGEL, M. 1998. *Dialogue acts in VERBMOBIL-2 (second edition)*. Verbmobil Report 226. Saarbrücken: DFKI.

ALLAN, K. 1998. *Meaning and speech acts*. Web published document, Linguistics Department, Monash University.

ALLEN, J. F. 1983. 'Recognising intentions from natural language utterances' in Brady, M. & Berwick, R. C. (Eds.): 107-166.

ALLEN, J. F. 1987. *Natural language understanding*. Menlo Park, Calif. Wokingham: Benjamin/Cummings Publishing Company.

ALLEN, J. F., SCHUBERT, L. K., FERGUSON, G., HEEMAN, P., HWANG, C. H., KATO, T., LIGHT, M., MARTIN, N. G., MILLER, B. W., POESIO, M & TRAUM, D. R. 1994. *The TRAINS project: A case study in defining a conversational planning agent*. Technical Report TR532, University of Rochester, Computer Science Department.

ALLWOOD, J. S. 1994. 'Obligations and options in dialogue', *Think*, **3** (1): 9-18.

ALLWOOD, J. S. 1995. 'An activity-based approach to pragmatics', *Gothenburg Papers in Theoretical Linguistics*, **76**.

ALLWOOD, J. S., NIVRE, J. & AHLSÉN, E. 1992. 'On the semantics and pragmatics of linguistic feedback', *Journal of Semantics*, **9:** 1-26.

ANSCOMBE, G. E. M. 1957. *Intention.* Oxford: Blackwell.

APPELT, D. E. 1985. *Planning English sentences.* Cambridge: Cambridge University Press.

ARDISSONO, L., BOELLA, G. & LESMO, L. 1998. 'Dialog modeling in an agent-based framework', *Proceedings of the 2nd International Workshop on Human -Computer Conversation*, Bellagio, Italy.

ARDISSONO, L., BOELLA, G. & LESMO, L. 2000. 'A plan based agent architecture for interpreting natural language dialogue', *International Journal of Human-Computer Studies* **52** (4): Academic Press.

ASHER, N. 1999. 'Discourse structure and the logic of conversation' in Turner, K. (Ed.).

ASHER, N. & LASCARIDES, A. 1994. 'Intentions and information in discourse', *Proceedings of the 29th Annual meeting of the Association of Computational Linguistics*, 34-41, Las Cruces, USA, June 1994.

AUSTIN, J. L. (FLEW, A. ED.) 1953. 'Other minds', *Logic and Language: Second Series.* Oxford: Basil Blackwell.

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

AUSTIN, J. L. 1971. 'Performative-constative' in Searle, J. R. (ed.).

AUSTIN, J. L. (URMSON, J. O. & SBISÀ, M. EDS.) 1975. *How to do things with words: The Wiliam James lectures delivered at Harvard University in 1955.* Oxford: Clarendon Press.

BACH, K. & HARNISH R. M. 1979. *Linguistic communication and speech acts.* Cambridge, Mass.: MIT Press.

BALL, G., LING, D., KURLANDER, D., MILLER, J. PUGH, D., SKELLY, T., STANKOSKY, A., THIEL, D., VAN DANTZICH, M. & WAX T. 1997. 'Lifelike computer characters: The Persona project at Microsoft', in Bradshaw (Ed.).

BALLMER, T. T. & BRENNENSTUHL, W. 1981. *Speech act classification: A study in the lexical analysis of English speech activity verbs.* Berlin Heidelberg New York: Springer-Verlag (Springer Series in Language and Communication, Vol. 8).

BERGER, P. L. & LUCKMANN, T. 1966. *The social construction of reality: A treatise in the sociology of knowledge.* Garden City, New York: Doubleday.

BERNSEN, N. O., DYBKJÆR, H. & DYBKJÆR, L. 1998. *Designing interactive speech systems: From first ideas to user testing.* London: Springer-Verlag.

BERNSEN, N. O., DYBKJÆR, L. & KOLODNYTSKY, M. 2002. 'THE NITE WORKBENCH – A tool for annotation of natural interactivity and multimodal data', *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'2002)*, Las Palmas, May 2002.

BERRY, M. 1981. 'Systemic linguistics and discourse analysis: A multi-layered approach to exchange structure' in Coulthard, M. & Montgomery, M. (Eds.).

BICKMORE, T., & CASSELL, J. 2001. 'Relational agents: A model and implementation of building user trust', *Proceedings of the Conference for Computer-Human Interaction*.

BOS, J., & GABSDIL, M. 2000. 'First-order inference and the interpretation of questions and answers', *Proceedings of Götalog 2000, 4th Workshop on the Semantics and Pragmatics of Dialogue*: 43-50.

BOYE, J., WIRÉN, M., RAYNER, M., LEWIN, I., CARTER, D. & BECKET, R. 1999. 'Language-processing strategies and mixed-initiative dialogues', *Proceedings of IJCAI-99 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.

BRADSHAW, J. (ED.) 1997. *Software agents.* AAAI/Cambridge: MIT Press.

BRADY, M., BERWICK, R. C., ALLEN, J., ET AL. (EDS.) 1983. *Computational models of discourse.* Cambridge, Mass. London: MIT Press.

BROWN, G. 1995. *Speakers, listeners, and communication: Explorations in discourse analysis.* Cambridge: Cambridge University Press.

BROWN, G. & YULE, G. 1983. *Discourse analysis.* Cambridge: Cambridge University Press.

BROWN, P. & LEVINSON, S. C. 1978. 'Universals in language usage: Politeness phenomena' in Goody, E. (Ed.).

BROWN, P. & LEVINSON, S. C. 1987. *Politeness: Some universals in language usage.* Cambridge: Cambridge University Press.

BUNT, H. 1989. 'Information dialogues as communicative action in relation to partner modelling and information processing' in Taylor, M. M., Bouwhuis, D. G. & Neel, F. (Eds.).

BUNT, H. 1994. 'Context and dialogue control', *Think*, **3** (1): 19-31.

BUNT, H. 1995. 'Dynamic interpretation in text and dialogue' in Taylor, M. M., Bouwhuis, D. G. & Neel, F. (Eds.).

BUNT, H. 2000. 'Dynamic interpretation and dialogue theory' in Taylor, M. M., Bouwhuis, D. G. & Neel, F. (Eds.).

BUNT, H. 2001. 'Dialogue pragmatics and context specification' in Bunt, H. & Black, W. (Eds.).

BUNT, H. & BLACK, W. (EDS.) 2001. *Abduction, belief and context in dialogue: Studies in computational pragmatics.* Amsterdam: John Benjamins.

CAMERON, D. 1992. 'Not gender difference but the difference gender makes: the politics of explaining sex differences in language' in Proceedings of the Tromsø Symposium on language and gender, Bull, T. and Swan, T. (Eds.), *International Journal of the Sociology of Language*, **94**:13-26

CARBERRY, S. 1990. *Plan recognition in natural language dialogue.* Cambridge, Mass.: MIT Press.

CARLETTA, J. 1996. 'Assessing agreement on classification tasks: The kappa statistic', *Computational Linguistics*, **22** (2): 249-254.

CARLETTA, J., ISARD, A., ISARD, S., KOWTKO, J. C., DOHERTY-SNEDDON, G. & ANDERSON, A. H. 1997. 'The reliability of a dialogue structure coding scheme', *Computational Linguistics*, **23** (1): 13-31.

CARLSON, L. 1983. *Dialogue games: An approach to discourse analysis.* Dordrecht London: Reidel.

CASSELL, J. 2000. 'More than just another pretty face: Embodied conversational interface agents', *Communications of the ACM*, **43** (4): 70-78.

CASSELL, J., TORRES, O. E., & PREVOST, S. 1999. 'Turn taking versus discourse structure', in Wilks (Ed.).

CASTELFRANCHI, C., DE ROSIS, F., & GRASSO, F. 1999. 'Deception and suspicion in medical interactions', in Wilks (Ed.).

CHANG, M. K. & WOO, C. C. 1994. 'A speech-act-based negotiation protocol: Design, implementation and test use', *ACM Transations on Information Systems*, **12** (4): 360-382.

CHANG, N. & FISCHER, I. 2000. 'Understanding idioms', *Konvens 2000 Sprachkommunikation*, ITG-Fachbericht **161**: 33-38.

CHOMSKY, N. 1965. *Aspects of the theory of syntaxl.* Cambridge, Mass.: MIT Press.

CHOMSKY, N., RIEBER, R. W. & VOYAT, G. 1983. *Dialogues on the psychology of language and thought: Conversations with Noam Chomsky ... [et al.].* New York: Plenum Press.

CHU-CARROLL, J. & BROWN, M. K. 1997. 'Tracking initiative in collaborative dialogue interactions', *Proceedings of ACL/EACL'97*.

CHURCHER, G. E. 1997. *Dialogue management in speech recognition applications.* Leeds: University of Leeds (PhD thesis).

CLARK, H. H. 1983. 'Making sense of nonce sense' in Flores d'Arcais, G. B. & Jarvella, R. J. (Eds.).

CLARK, H. H. 1996. *Using language.* Cambridge: Cambridge University Press.

CLARK, H. H. & CARLSON, T. B. 1982. 'Hearers and speech acts', *Language*, **58**: 332-373.

CLARK, H. H. & SCHAEFER, E. F. 1989. 'Contributing to discourse', *Cognitive Science*, **13**: 259-294.

COHEN, P. R. & LEVESQUE, H. J. 1990. 'Intention is choice with commitment', *Artificial Intelligence*, **42**: 213-261.

COHEN, P. R., MORGAN, J. L. & POLLACK, M. E. 1990. *Intentions in communication.* Cambridge, Mass. London: MIT Press.

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

COHEN, P. R., PERRAULT, C. R. & ALLEN, J. F. 1982. 'Beyond question answering' in Lehnert, W. & Ringle, M. (Eds.).

COLBY, K. M. 1973. 'Simulations of belief systems', in Schank & Colby (Eds.).

COLE, P. (ED.) 1978. *Syntax and semantics: Vol. 9 Pragmatics.* New York London: Academic Press.

COLE, P. & MORGAN, J. L. (EDS.) 1975. *Syntax and semantics: Vol. 3 Speech acts.* New York London: Academic Press.

COOPER, R. & LARSSON, S. 1999. 'Dialogue moves and information states', *Proceedings of the 3rd International Workshop on Computational Semantics.*

CORE, M. G. 1998. 'Analyszing and predicating patterns of DAMSL utterance tags', AAAI Spring Symposium on Applying Machine Learning to Discourse Processing, Stanford, CA.

CORE, M. G. & ALLEN, J. F. 1997a. 'Coding dialogs with the DAMSL annotation scheme', AAAI Fall Symposium on Communicative Action in Humans and Machines, Boston, MA.

CORE, M. G. & ALLEN, J. F. 1997b. *Draft of DAMSL: Dialog act markup in several layers*. Coder's Manual (unpublished).

COULTHARD, R. M. AND BRAZIL D. 1979. *Exchange structure: Discourse analysis monographs no. 5.* Birmingham: The University of Birmingham, English Language Research.

DAVIDSON, D. 1974. 'On the very idea of a conceptual scheme', *Proceedings and Addresses of the American Philosophical Association*: 5-20.

DAVIDSON, D. & HARMAN, G. H. (EDS.) 1972. *Semantics of natural language.* Dordrecht: Reidel.

DOWNEY, S., BREEN, A. P., FERNANDEZ, M. & KANEEN, E. 1998. 'Overview of the Maya Spoken Language System', *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98).* .

EGGINS, S. & SLADE, D. 1997. *Analysing casual conversation.* London New York: Cassell.

ESA, A. & LYYTINEN, K. 1996. 'On the success of speech acts and negotiating commitments', *Proceedings of the 1st International Workshop on Communication Modelling*, Tilburg, The Netherlands (Springer-Verlag).

FERGUSON, G., ALLEN, J. F. & MILLER, B. 1996. 'TRAINS-95: Towards a mixed-initiative planning assistant' in *Proceedings of the 3rd Conference on Artificial Intelligence Planning Systems (AIPS-96)*, Edinburgh, Scotland: 70-77.

FIKES, R. E. & NILSSON, N. J. 1971. 'STRIPS: A new approach to the application of theorem proving to problem solving', *Artificial Intelligence*, **2** (3/4): 189-208.

FININ, T., LABROU, Y. & MAYFIELD, J. 1997. 'KQML as an agent communication language' in Bradshaw (Ed.).

FIRTH, J.R. 1957. 'A synopsis of linguistic theory, 1930-1955' in Palmer, F.R. (Ed.).

FLORES D'ARCAIS, G. B. & JARVELLA, R. J. (EDS.) 1983. *The process of language understanding.* Chichester: Wiley.

FODOR, J. A. 1987. *Psychosemantics: The problem of meaning in the philosophy of mind.* Cambridge, Mass.; London: MIT Press.

FODOR, J. A. 1990. 'Banish discontent' in Lycan, W. G. (Ed.), (first published in 1986).

FODOR, J. D. 1983. *The modularity of mind*. Cambridge, Mass.: MIT Press.

FUKADA, T., KOLL, D., WAIBEL, A. & TANIGAKI, K. 1998. 'Probabilistic dialogue act extraction for concept based multilingual translation systems', *ICSLP-98*, Sydney, Australia.

GAZDAR, G. 1981. 'Speech act assignment' in Joshi, A. K. et al (Eds.).

GIBBS, R. W. 1994. *The poetics of mind: Figurative thought, language and understanding.* Cambridge: Cambridge University Press.

GIBBS, R. W. 1996. 'Skating on thin ice: Literal meaning and understanding idioms in conversation', *Discourse Processes*, **9**: 17-30.

GIBBS, R. W. 2002. 'A new look at literal meaning in understanding what is said and implicated', *Journal of Pragmatics*, **34**: 457-486.

GINZBURG, J. 1996. 'Interrogatives: Questions, facts and dialogue', in Lapin (Ed.).

GIORA, R. & FEIN, O. 1999. 'Irony: Context and salience', *Metaphor and Symbol*, **14**: 241-257.

GLUCKSBERG, S. 1991. 'Beyond literal meaning: The psychology of allusion', *Current Directions in Psychological Science*, **2**: 146-152.

GLUCKSBERG, S. 1998. 'Metaphor', *Current Directions in Psychological Science*, **7**: 39-43.

GOFFMAN, E. (ED.) 1967. *Interaction ritual: Essays on face-to-face behaviour.* London: Allen Lane, The Penguin Press.

GOFFMAN, E. 1976. 'Replies and responses', *Language in Society*, **5**: 257-313.

GOLDBERG, A. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

GOODWIN, C. 1981. *Conversational organization: Interaction between speakers and hearers.* New York London: Academic Press.

GOODY, E. N. (ED.) 1978. *Questions and politeness: Strategies in social interaction.* Cambridge: Cambridge University Press.

GORDON, D. & LAKOFF, G. 1975. 'Conversational postulates', in Cole & Morgan (Eds.).

GRAU, B., SABAH, G. & VILNAT, A. 1994. 'Control in man-machine dialogue', *Think*, **3** (1): 32-55.

GRICE, H. P. 1957. 'Meaning', *Philosophical Review*, **66**: 377-388.

GRICE, H. P. 1968. 'Utterer's meaning, sentence-meaning and word-meaning', *Foundations of Language*, **4** (1968): 1-18 (also in Searle 1971: 54-70).

GRICE, H. P. 1975. 'Logic and conversation' in Cole, P. & Morgan, J. L. (Eds.).

GRICE, H. P. 1978. 'Further notes on logic and conversation' in Cole, P. (Ed.).

GROSZ, B. J. 1977. *The representation and use of focus in dialogue understanding*. Technical report, Artificial Intelligence Center, SRI International, Menlo Park, C.A.

GROSZ, B. J. 1978a. 'Focusing in dialog' in *Proceedings of the Second Workshop on Theoretical Issues in Natural Language Processing (TINLAP-2)*, 96-103, University of Illinois, Urban-Champaign, July.

GROSZ, B. J. 1978b. 'Understanding spoken language' in Walker, D. (Ed.).

GROSZ, B. J. 1981. 'Focusing and description in natural language dialogs' in Joshi, A. K., Webber, B. L. and Sag, I. A. (Eds.).

GROSZ, B. J. & KRAUS, S. 1999. 'The evolution of shared plans' in Rao, A. & Wooldridge, M. (Eds.): 227-262.

GROSZ, B.J., & SIDNER, C.L. 1986. 'Attention, intentions, and the structure of discourse', *Computational Linguistics*, **12** (3): 175-204.

GROSZ, B.J., & SIDNER, C.L. 1990. 'Plans for discourse' in Cohen, P. R., Morgan, J. L. & Pollack, M. E. (Eds.).

GUMPERZ, J. 1982. *Discourse strategies.* Cambridge: Cambridge University Press.

GUMPERZ, J. & HYMES, D. 1972. *Directions in sociolinguistics: The ethnography of communication.* New York: Holt, Rinehart & Winston.

HAGEN, E. & POPOWICH, F. 2000. 'Flexible speech act based dialogue management', *Proceedings of the 1st Sigdial Workshop at ACL2000.*

HALLIDAY, M. A. K. 1961. 'Categories of the Theory of Grammar', *Word*, **17** (3): 241-292.

HALLIDAY, M. A. K. 1973. *Explorations in the functions of language.* London: Edward Arnold.

HALLIDAY, M. A. K. 1975. *Learning how to mean*. London: Edward Arnold.

HALLIDAY, M. A. K. 1994. *An introduction to functional grammar.* London: Edward Arnold.

HAMBLIN, C. L. 1971. 'Mathematical models of dialogue', *Theoria*, **37**: 130-155.

HAMBLIN, C. L. 1970. *Fallacies.* London: Methuen.

HASSELL, L. & CHRISTENESEN, M. 1996. 'Indirect speech acts and their use in three channels of communication', *Proceedings of the 1st International Workshop on Communication Modelling*, Tilburg, The Netherlands (Springer-Verlag).

HEAP, A. & HOGG, D. 1996. 'Toward 3D hand tracking using a deformable model', *Proceedings of the 2nd International Conference on Face and Gesture Recognition*.

HEDENIUS, I. 1963. 'Performatives', *Theoria*, **29**: 115-136.

HERINGER, J. T. 1972. *Some grammatical correlates of felicity conditions and presuppositions.* Mimeo: Indiana University Linguistics Club.

HIRSCHBERG, J. 1989. 'Distinguishing questions by contour in speech recognition tasks', *Proceedings of the DARPA Speech and Natural Language Processing Workshop*, Cape Cod MA, October. San Francisco: Morgan Kauffman.

HIRSCHBERG, J. 2000. 'A Corpus-Based Approach to the Study of Speaking Style', in Horne (Ed.).

HINKELMAN, E. A. 1989. *Linguistic and pragmatic constraints on utterance interpretation*. Ph.D. Thesis, Computer Science Department, University of Rochester, N.Y.

HINKELMAN, E. A. & ALLEN, J. F. 1989. 'Two constraints of speech act interpretation', *Proceedings of the Association for Computational Linguistics*: 212-219.

HOLDCROFT, D. 1978. *Words and deeds: Problems in the theory of speech acts.* Oxford: Clarendon Press.

HOOK, S. (ED.) 1969. *Language and philosophy: A symposium.* New York: New York University Press.

HORNE, M. (ED.) 2000. *Festschrift in Honor of Gösta Bruce*. Amsterdam: Kluwer.

HOUGHTON, G. 1986. *The production of language in dialogue*. Ph.D. Thesis, University of Sussex.

HOUSE, D., BESKOW, J. & GRANSTRÖM, B. 2001. 'Interaction of visual cues for prominence', *Working Papers*, **49**: 62-65.

HULSTIJN, J. 2000. 'Dialogue games are recipes for joint action', *Proceedings of the 4th Workshop on the Semantics and Pragmatics of Dialogue (GÖTALOG 2000)*.

HUMBERSTONE, L. 1992. 'Direction of fit', *Mind,* January.

HYMES, D. 1972. 'Models of the interaction of language and social life' in Gumperz, J. and Hymes, D. (Eds.): 35-71.

JACOBS, R. & ROSENBAUM, P. (EDS.) 1970. *Readings in English transformation grammar*. Waltham, MA: Ginn.

JEFFERSON, G. 1974. 'Error-correction as an interactional discourse', *Language in Society*, **3**:181-200.

JEFFERSON, G. 1978. 'Sequential aspects of storytelling in conversation', *Studies in the Organization of Conversational Interaction*. 219-248. Academic Press.

JEKAT, S., KLEIN, A., MAIER, E., MALECK, I. MAST, M., & QUANTZ, J. 1995. *Dialogue acts in VERBMOBIL*. Technical Report DFKI Saarbrüchen, VM-Report 65.

JOHNSON-LAIRD, P. N. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge: Cambridge University Press.

JOHNSON-LAIRD, P. N. 1990. 'What is communication?' in Mellor, D. H. (Ed.): 1-14.

JOSHI, A. K., WEBBER, B. L. & SAG, I. A. (EDS.) 1981. *Elements of discourse understanding.* Cambridge: Cambridge University Press.

JURAFSKY, D., SHRIBERG, E. & BIASCA, D. 1997a. *SWITCHBOARD SWBD-DAMSL shallow-discourse-function annotation.* Coder's Manual, Draft 13, University of Colorado at Boulder and SRI International.

JURAFSKY, D., BATES, R., COCCARO, N., MARTIN, R., METEER, M., RIES, K., SHRIBERG, E., STOLKE, A., TAYLOR, P. & ESS-DYKEMA, C. VAN. 1997b. *SWITCHBOARD discourse language modeling project.* Johns Hopkins LVCSR Workshop-97.

KAMP, H. & REYLE, U. 1993. *From discourse to logic.* Studies in Linguistics and Philosophy, **42**. Dordrecht, Boston, London: Kluwer Academic Press.

KATZ, J. J. 1972. *Semantic theory.* New York London: Harper and Row.

KATZ, J. J. 1977. *Propositional structure and illocutionary force: A study of the contribution of sentence meaning to speech acts.* Hassocks: Harvester Press.

KEARNS, J. T. 1984. *Using language: The structures of speech acts.* Albany: State University of New York Press.

KEENAN, E. L. (ED.) 1975. *Formal semantics of natural language: Papers from a colloquium sponsored by the King's College Research Centre, Cambridge.* Cambridge: Cambridge University Press.

KEIZER, S. 2001. '*A Bayesian approach to dialogue act classification', Proceedings of the 5th Workshop on the Semantics and Pragmatics of Dialogue (BI-DIALOG 2001)*, University of Bielefeld, Germany.

KIPP, M., REITHINGER, N., BERNSEN, N. O., DYBKJÆR, L., KNUDSEN, M. W., MACHUCA, M. & RIERA, M. 2002. Best practice gesture, facial expression, and cross-modality coding schemes for inclusion in the workbench. NITE Deliverable D2.3.

KITA, K., YOSHIKANZU, F., MASAAKI, N. & TSUYOSHI, M. 1996. 'Automatic acquisition of probabilistic dialogue models', *ICSLP-96*, Philadelphia: 196-199.

KLEIN, M., BERNSEN, N.O., DAVIES, S., DYBKJÆR, L., GARRIDO, J., KASCH, H., MENGEL, A., PIRRELLI, V., POESIO, M., QUAZZA, S. & SORIA, C. 1998. *Supported coding schemes.* MATE Deliverable D1.1, LE Telematics Project LE4-8370.

KOWTKO, J. C., ISARD, S. D. & DOHERTY, G.M. 1992. *Conversational games within dialogue*. HCRC Research Paper RP-31, Edinburgh.

KRIPPENDORF, K. 1980. *Content Analysis*. Sage Publications.

LABOV, W. 1970. 'The study of language in its social context', *Studium Generale,* **23**: 30-87.

LABOV, W. & FANSHEL, D. 1977. *Therapeutic discourse: Psychotherapy as conversation.* New York: Academic Press.

LAFFAL, J. 1965. *Pathological and normal language.* New York: Atherton Press.

LAKOFF, G. 1972. 'Linguistics and natural logic' in Davidson & Harman (Eds.).

LAKOFF, G. 1975. 'Pragmatics and natural logic' in Keenan (Ed.).

LAKOFF, G. 1987. *Woman, fire and dangerous things: What categories reveal about the mind*. University of Chicago Press.

LAPIN, S. (ED.) 1996. *Handbook of Contemporary Semantic Theory*. Oxford: Blackwell.

LAURENCEAU, J., BARRETT, L. & PIETROMONACO, P. 1998. 'Intimacy as an interpersonal process: the importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges', *Journal of Personality and Social Psychology*, **74**: (5) 1238-1251.

LEE, M. 1997. 'Belief ascription in mixed initiative dialogue', *Proceedings of AAAI Spring Symposium on Mixed Initiative Interaction*.

LEE, M. & WILKS, Y. 1996. 'An ascription-based approach to speech acts', *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, Denmark.

LEE, M. & WILKS, Y. 1997. 'Eliminating deceptions and mistaken belief to infer conversational implicature', *Proceedings of IJCAI-97 Workshop on Conflict, Co-operation and Collaboration in Dialogue Systems*.

LEECH, G. N. 1980. *Language and tact (Pragmatics and beyond series)*. Amsterdam: John Benjamins.

LEECH, G. N. 1983. *Principles of pragmatics.* London: Longman.

LEHNERT, W. & RINGLE, M. (EDS.) 1982. *Strategies for Natural Language Processing*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

LERNER, G. H. 1987. *Collaborative turn sequences: sentence construction and social action*. Irvine: University of California, Ph.D. dissertation.

LEVIN, J. A. & MOORE, J. A 1978. 'Dialogue-games: Metacommunication structures for natural language interaction', *Cognitive Science*, **1** (4): 395-420.

LEVINSON, S. C. 1983. *Pragmatics.* Cambridge: Cambridge University Press.

LEWIN, I., RUSSELL, M., CARTER, D., BROWNING, S., PONTING, K. & PULMAN, S. G. 1993. 'A speech-based route enquiry system built from general-purpose components', *Proceedings of the 3rd European Conference on Speech Communication and Technology (EUROSPEECH '93)*, Vol. 3: 2047-2050.

LEWIN, I. 2000. 'A formal model of conversational game theory', *Proceedings of the 4th Workshop on the Semantics and Pragmatics of Dialogue (GÖTALOG 2000)*.

LITMAN, D. L. 1985. *Plan recognition and discourse analysis: An integrated approach for understanding dialogues*. Ph.D. Dissertation and Technical Report TR-170, University of Rochester, N.Y.

LITMAN, D. L. & ALLEN, J. F. 1984. 'A plan recognition model for clarification subdialogues', *Proceedings of COLING-94*, 302-311.

LITMAN, D. L. & ALLEN, J. F. 1987. 'A plan recognition model for subdialogues in conversation', *Cognitive Science*, **11**: 163-200.

LITMAN, D. L. & ALLEN, J. F. 1990. 'Discourse processing and commonsense plans' in Cohen, P. R., Morgan, J. L. & Pollack, M. E. (Eds.).

LOCHBAUM, K. E. 1994. *Using collaborative plans to model the intentional structure of discourse.* Cambridge, Mass.: Harvard University (Ph. D. dissertation).

LOCHBAUM, K. E., GROSZ, B. J. & SIDNER, C. 2000. 'Discourse structure and intention recognition' in Dale, R., Moisl, H. & Somers, H. (Eds.).

LOCKE, J. 1971. *An essay concerning human understanding.* London, Glasgow: Fontana (first printed 1689).

LYCAN, W. G. (ED.) 1990. Mind and cognition. Oxford: Basil Blackwell (420-39).

LYONS, J. 1977. *Semantics (Vols. 1 & 2)*. Cambridge: Cambridge University Press.

MANN, W. C. 1988. 'Dialogue games: Conventions of human interaction', *Argumentation*, 2: 511-532.

MANN, W. C., MOORE, M., LEVIN, J. & CARLISLE, J. 1975. *Observation methods for human dialogue.* Technical Report RR/75/33, Information Sciences Institute, Marina del Rey, C. A.

MANN, W. C. & THOMPSON, S. 1983. *Relational propositions in discourse.* Technical Report RR/83/115, Information Sciences Institute, Marina del Rey, C. A.

MARCU, D. 2000. 'Perlocutions: The Achilles' heel of speech act theory', *Journal of Pragmatics*, **32**: 1719-1741.

MASSARO, D. W. & STORK, D. G. 1998. 'Speech recognition and sensory integration: A 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech', *American Scientist*, **86** (3): 236-244.

MAST, M., KOMPE, R., HARBECK, S., KIEßLING, A., NIEMANN, H. & NÖTH, E. 1996. 'Dialog act classification with the help of prosody', ICSLP-96, Philadelphia: 1728-1731.

MCCARTHY, J. 1990. *Elephant 2000 – A programming language based on speech acts.* Unpublished (deposited with CogPrints – Cognitive Sciences Eprint Archive in 1998)

MCDONALD, S. & PEARCE, S. 1996. 'Clinical insights into pragmatic language theory: The case of sarcasm', *Brain and Language*, **53**: 81-104.

MCGURK, H. & MACDONALD, J. 1976. 'Hearing lips and seeing voices', *Nature*, **264**: 746-748.

MCKEVITT, P.,PARTRIDGE, D. & WILKS, Y. 1992. 'Approaches to natural language discourse processing', *AI Review*, **6** (4): 333-364.

MCROY, S. W. & HIRST, G. 1995. 'The repair of speech act misunderstanding by abductive inference', *Computational Linguistics*, **21** (4): 435-478.

MCTEAR, M. F. (Submitted). *Spoken dialogue technology: Enabling the conversational user interface*. ACM Computing Surveys.

MELLOR, D. H. (ED.) 1990. *Ways of communicating.* Cambridge: Cambridge University Press.

MENGEL, A., DYBKJÆR, L., GARRIDO, J., HEID, U., KLEIN, M., PIRRELLI, V., POESIO, M., QUAZZA, S., SCHIFFRIN, A. & SORIA, C. 2000. *MATE dialogue annotation guidelines.* MATE Deliverable D2.1, LE Telematics Project LE4-8370.

NAGATA, M. & MORIMOTO, T. 1994. 'First steps toward statistical modeling of dialogue to predict the speech act type of the next utterance', *Speech Communication*, **15**: 193-203.

NARAYANAN, S. 1997. *Knowledge-based Action Representations for Metaphor and Aspect (KARMA)*. University of Berkeley dissertation.

PALMER, F. R. (ED.) 1968. *Selected papers of J.R. Firth 1952-9.* Harlow: Longman.

PARSONS, S. D. & JENNINGS, N. R. 1996. 'Negotiation through argumentation - a preliminary report', *Proceedings of the 2nd International Workshop on Multi-Agent Systems*, Kyoto, Japan, 1996: 267-274.

PIAGET, J. 1959. *The language and thought of the child.* London: Routledge & Kegan Paul.

POESIO, M. & TRAUM, D. 1997. 'Conversational actions and discourse situations', *Computational Intelligence*, **13**(3): 309-347.

POESIO, M. & TRAUM, D. 1998. 'Towards an axiomatization of dialogue acts', *Proceedings of the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues (13th Twente Workshop on Language Technology)*, J. Hulstijn and A. Nijholt (Eds.), Enschede, 207-222, May 1998.

POLLACK, M. E. 1986. *Inferring domain plans in question-answering.* Ph.D. Dissertation, University of Pennsylvania, Philadelphia.

POWER, R. 1979. 'The organization of purposeful dialogs', *Linguistics*, **17**: 105-152.

PULMAN, S. G. 1996. 'Conversational games, Belief Revision and Bayesian Networks', *Proceedings of the 7th Computational Linguistics in the Netherlands Meeting (CLIN-VII)*.

QUINE, W. V. O. 1960. *Word and object.* Cambridge, Mass.: Technology Press of the Massachusetts Institute of Technology London: Wiley.

RAE, J. P. 1989. *Explanations and communicative constraints in naturally occurring discourse.* Leeds: University of Leeds (PhD thesis).

RAO, A. & WOOLDRIDGE, M. (EDS.) 1999. *Foundations and theories of rational agency.* Kluwer Academic Press.

RAWLS, J. 1955. 'Two concepts of rules', *Philosophical Review*, **64**: 3-13.

REICHMAN, R. 1985. *Getting computers to talk like you and me: discourse context, focus and semantics (an ATN model)*. Cambridge, Mass. London : MIT Press.

REICHMAN-ADAR, R. 1984. 'Extended person-machine interfaces', *Artificial Intelligence*, **22**: 157-218.

REITHINGER, N., ENGEL, R., KIPP, M. & KLESEN, M. 1996. 'Predicting dialgogue acts for a speech-to-speech translation system', *ICSLP-96*, Philadelphia: 654-657.

REITHINGER, N. & KLESEN, M. 1997. 'Dialogue act classification using language models', *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, Rhodes, Greece.

ROSS, J. R. 1970. 'On declarative sentences' in Jacobs, R. & Rosenbaum, P. (Eds.). 222-272.

RUMELHART, D. E., MCCLELLAND, J. L. ET AL. 1986. *Parallel distributed processing: Explorationsin the microstructure of cognition*. Cambridge, Mass.: MIT Press.

SACERDOTI, E. 1977. *A structure for plans and behaviour.* Amsterdam: North-Holland.

SACKS, H., SCHEGLOFF, E. A. & JEFFERSON, G. 1974. 'A simplest systematics for the organisation of turn-taking in conversation', *Language*, **50** (4): 696-735.

SADOCK, J. M. 1974. *Toward a linguistic theory of speech acts.* New York London: Academic Press.

SAUSSURE, F. DE 1916. *Cours de linguistique générale*. Paris: Payot.

SBISÀ, M. 2001. 'Illocutionary force and degrees of strength in language use', *Journal of Pragmatics*, **33**: 1791-1814.

SCHANK, R. C. & COLBY, K. M. (EDS.) 1973. *Computer models of thought and language.* San Francisco: W. H. Freeman.

SCHEGLOFF, E. A. 1968. 'Sequencing in conversational openings', *American Anthropologist*, **70**: 1075-1095.

SCHEGLOFF, E. A. 1976. *On some questions and ambiguities in conversation.* Pragmatics Microfiche, 2.2: D8-G12.

SCHEGLOFF, E. A. 1988. 'Presequences and indirection', *Journal of Pragmatics*, **12**: 55-62.

SCHEGLOFF, E. A., JEFFERSON, G. & SACKS, H. (EDS.) 1977. 'The preference for self-correction in the organisation of repair in conversation', *Language*, **53**: 361-382.

SCHEGLOFF, E. A. & SACKS, H. 1973. 'Opening up closings', *Semiotica*, **7** (4): 289-327.

SCHIFFRIN, A. 1995. *A computational treatment of conversational contexts in the theory of speech acts.* Leeds: The University of Leeds (M.Sc. dissertation).

SCHIFFRIN, A. & MILLICAN, P. J. R. 1998. 'Identifying speech acts in context', *Proceedings of the 2nd International Workshop on Human -Computer Conversation*, Bellagio, Italy.

SCHIFFRIN, A. & SOUTER, D. C. 2002. 'Evaluating a Pragmatic Model of Speech Act Assignment Using Spoken Language Corpora', *Paper presented at the 23nd International Conference on English Language Research on Computerized Corpora of Modern and Medieval English (ICAME)*, Göteborg, 22-26 May 2002.

SCHIFFRIN, D. 1987. *Discourse markers.* Cambridge: Cambridge University Press.

SCHIFFRIN, D. 1994. *Approaches to discourse.* Oxford: Blackwell.

SEARLE, J. R. 1969. *Speech acts: an essay in the philosophy of language.* London: Cambridge University Press.

SEARLE, J. R. (ED.) 1971. *The philosophy of language.* London: Oxford University Press.

SEARLE, J. R. 1975. 'Indirect speech acts' in Cole, P. & Morgan, J. L. (Eds.).

SEARLE, J. R. 1979. *Expression and meaning: Studies in the theory of speech acts.* Cambridge: Cambridge University Press.

SEARLE, J. R. 1983. *Intentionality: an essay in the philosophy of mind.* Cambridge: Cambridge University Press.

SEARLE, J. R. & VANDERVEKAN, D. 1985. *Foundations of illocutionary logic.* Cambridge: Cambridge University Press.

SERENARI, M., DYBKJÆR, L., HEID, U., KIPP, M. & REITHINGER N. 2002. *Survey of Existing Gesture, Facial Expression, and Cross-modality Coding Schemes*. NITE Deliverable D2.1.

SHANAHAN, M. P. 1999. 'The event calculus explained', *Artificial Intelligence Today*, Wooldridge, M.J. & Veloso, M. (Eds.), Springer Lecture Notes in Artificial Intelligence No. **1600**, Springer: 409-430.

SIERRA, C., JENNINGS, N. R., NORIEGA, P. & PARSONS, S. 1997. 'A framework for argumentation-based negotiation', *Proceedings of the 4th International Workshop on Agent Theories, Architectures and Languages (ATAL-97)*, Rhode Island, USA.

SIDNER, C. L. 1983. 'What the speaker means: The recognition of speakers' plans in discourse', *International Journal of Computers and Mathematics, Special Issue in Computational Linguistics*, **9** (1): 71-82.

SIDNER, C. L. 1985. 'Plan parsing for intended response recognition in discourse', *Computational Intelligence*, **1** (1): 1-10.

SINCLAIR, J. M. & COULTHARD, M. 1975. *Towards an analysis of discourse: The English used by teachers and pupils.* London: Oxford University Press.

SLOBIN, D. I. 1975. 'The more it changes… on understanding language by watching it move through time', *Papers and Reports on Child Language Development*, University of California, Berkeley, September 1975: 1-30.

SMITH, P. W. H. 1991. *Speech act theory, discourse structure and indirect speech acts.* Leeds: University of Leeds (PhD thesis).

SMITH, R. W. & GORDON, S. A. 1997. 'Effects of variable initiative on linguistic behaviour in human-computer spoken natural language dialogue', *Computational Linguistics*, **23** (1): 141-168.

SPERBER, D. & WILSON, D. 1986. *Relevance.* Oxford: Blackwell.

SPERBER, D. & WILSON, D. 1995. *Relevance: Communication and cognition.* Oxford: Blackwell.

STARNER, T. & PENTLAND, A. 1996. *Real-time American Sign Language recognition from video using hidden Markov models*, Technical report, MIT Medialab, Vision and Modeling Group, TR-375.

STENIUS, E. 1967. 'Mood and language game', *Synthese*, 1967: 254-274.

STENSTRÖM, A. 1995. *An introduction to spoken interaction (learning about language)*. London: Longman.

STOLCKE, A., COCCARO, N., BATES, R., TAYLOR, P., VAN ESS-DYKEMA, C., RIES, K., SHRIBERG, E., JURAFSKY, D., MARTIN, R. & METEER, M. 2000. 'Dialogue act modeling for automatice tagging and recognition of conversational speech', *Computational Linguistics*, **26** (3): 339-373.

SUHM, B. & WAIBEL, A. 1994. 'Toward better language models for spontaneous speech', ICSLP-94: 831-834.

TANAKA, H. & YOKOO, A. 1999. 'An efficient statistical speech act type tagging system for speech translation systems', *Proceedings of the 37th Conference of the Association for Computational Linguistics*.

TANNEN, D. 1984. *Conversational style: Analyzing talk among friends.* Norwood, N.J.: Ablex Publishing Corporation.

TANNEN, D. 1989. *Talking voices: Repetition, dialogue, and imagery in conversational discourse.* Cambridge: Cambridge University Press.

TAYLOR, M. M., BOUWHUIS, D. G. & NEEL, F. (EDS.) 1989. *The structure of multimodal dialogue.* Amsterdam: Elsevier Science Publishers.

TAYLOR, M. M., BOUWHUIS, D. G. & NEEL, F. (EDS.) 1995. *The structure of multimodal dialogue.* Vol. 2, Amsterdam: John Benjamins.

TAYLOR, M. M., BOUWHUIS, D. G. & NEEL, F. (EDS.) 2000. *The structure of multimodal dialogue.* Vol. 2, Amsterdam: John Benjamins.

TAYLOR, P., KING, S., ISARD, S. & WRIGHT, H. 1998. 'Intonation and dialogue context as constraints for speech recognition', *Language and Speech*, **41** (3-4): 489-508.

TAYLOR, T. J. & CAMERON, D. 1987. *Analysing conversation: Rules and units in the structure of talk.* Oxford: Pergamon.

TEMPLE, J. & HONECK, R. 1999. 'Proverb comprehension: The primacy of literal meaning', *Journal of Psycholinguistic Research*, **28**: 41-70.

THOMAS, J. 1986. 'Role switching in interaction', *Lancaster Papers in Linguistics*, **34**: 1-19.

THOMAS, J. & SHORT, M. H. (EDS.) 1996. *Using corpora for language research: Studies in honour of Geoffrey Leech.* London: Longman.

TRAUM, D. R. & HINKELMAN, E. A. 1992. *Conversation acts in task-oriented spoken dialogue.* Technical Report TR425, University of Rochester, Computer Science Department.

TRINDI. 2001. *The TRINDI book: Task oriented instructional dialogue*. Technical Report LE4-8314, Gothenburg University, Sweden.

TSUI, A. M. B. 1994. *English conversation.* Describing English Language Series, London: Oxford University Press.

TSUI, A. M. B. 1995. *Introducing classroom interaction.* London: Penguin.

TURNER, K. (ED.) 1999. *The semantics/pragmatics interface from different points of view.* Oxford: Elsevier Science.

URMSON, J. O. & WARNOCK, G. J. (EDS.) 1979. *Philosophical papers / J.L. Austin.* Oxford: Oxford University Press.

VAISSIÈRE, J. 1995. 'Phonetic Explanations for Cross-Linguistic Prosodic Similarities', *Phonetica*, **52**: 123-130.

VYGOTSKII, L. S. (TRANS. & ED. BY HANFMANN, E. & VAKAR, G.) 1962. *Thought and language.* Cambridge,Mass.: M.I.T. Press.

WALKER, D. (ED.) 1978. *Discourse analysis*. New York: Elsevier/North-Holland.

WALLIS, H. L. B. 1994. 'Modelling argumentation using commitments', *Proceedings of the 3rd International Conference on Argumentation*, Amsterdam.

WALLIS, H. L. B. 1995. *Mood, speech acts and context.* Leeds: University of Leeds (Ph.D. thesis).

WANG, Y. & WAIBEL, A. 1997. 'Statistical analysis of dialogue structure', *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, Rhodes, Greece.

WATSON, R. 1993. *A survey of gesture recognition techniques.* Technical report, Department of Computer Science, Trinity College, Dublin, TCD-CS-93-11.

WIDDOWSON, H. G. 1978. *Teaching language as communication.* Oxford: Oxford University Press.

WILKS, Y. A. (ED.) 1999. *Machine conversations*. Kluwer.

WILLIAMS, S. 1996. 'Dialogue management in a mixed-initiative, cooperative spoken language system', *Proceedings of the 11th Twente Workshop on Language Technology (TWLT11)*, Enschade, Netherlands, 1996: 199-208.

WINOGRAD, T. 1983. *Language as a cognitive process.* Reading, Mass. London: Addison-Wesley.

WEIZENBAUM, J. 1976. *Computer power and human reason.* San Francisco: W. H. Freeman.

WITTGENSTEIN, L. 1953. *Philosophical investigations.* Oxford: Blackwell.

WOOLDRIDGE, M. & JENNINGS, N. R. 1994a. 'Towards a theory of cooperative problem solving', *Proceedings of the 6th European Workshop on Modelling Autonomous Agents in Multi-Agent Worlds (MAAMAW-94)*.

WOOLDRIDGE, M. & JENNINGS, N. R. 1994b. 'Formalizing the cooperative problem solving process', Klein, M. (Ed.) *Proceedings of the Thirteenth International Workshop on Distributed Artificial Intelligence (IWDAI-94)*, Lake Quinalt, WA, July 1994.

WOSZCZYNA, M. & WAIBEL, A. 1994. 'Inferring linguistic structure in spoken language', *Proceedings of the 1st International Conference on Spoken Language Processing (ICSLP-94)*, Yokohama, Japan: 847-850.

YALE, T. W. 1998. 'Integrating MindNet with HAL', *Proceedings of the 2nd International Workshop on Human-Computer Conversation*, Bellagio, Italy.

ZIFF, P. 1969. 'Natural and formal language' in Hook, S. (Ed.).