The production of prosodic focus and contour in dialogue

Thesis

The production of prosodic focus and contour in dialogue

Nicholas John Youd

BA, MSc

Thesis submitted for the degree of PhD

The Open University

Relevant disciplines: Computational Linguistics, Cognitive Science

16th December 1992

Abstract

Computer programs designed to converse with humans in natural language provide a framework against which to test supra-sentential theories of language production and interpretation. This thesis seeks to flesh out, in terms of a computer model, two basic assumptions concerning prosody—that speakers use intonation to convey intention, or attitude, and that prosodic prominence serves to convey conceptual prominence.

A model of an information-providing agent in is proposed, based on an analysis of a corpus of spontaneous dialogues. This uses an architecture of communicating processes, which perform interpretation, application-specific planning, repair, and the production of output. Dialogue acts are then defined as feature bundles corresponding to significant events. A corpus of read dialogues is analysed in terms of these features, and using conventional intonational labelling. Correlations between the two are examined.

Prosodic prominence is examined at three levels. At the level of surface encoding, re-use of substrings and structural parallelism can reduce processing for the speaker, and the listener. At the level of conceptual planning, similar benefits exist, given that speakers and listeners assume a common discourse model wherever possible. At these levels use is made of a short-term buffer of recent forms. A speaker may additionally use contrastive prominence to draw the listener's attention to disparities. Finally, at the level of intentions, a speaker wish to highlight certain information, regardless of accessibility.

Prosodic focus is represented relationally, rather than via a simple binary-valued feature. This has the advantage of facilitating the mapping between levels; it also renders straightforward the notion of focus as the product of a number of potentially conflicting influences.

Those parts of the theory concerned with discourse representation, language generation, and prosodic focus have been implemented as part of the Sundial dialogue system. In this system, discoursal and pragmatic decisions affecting prosody are converted to annotations on a text string, for realisation by a rule-based synthesizer.

Contents

List of Figures

List of Tables

Acknowledgements

Thanks to my internal supervisor, George Kiss, for playing an enabling role, and for advice and discussions, particularly during the early phases of my research.

During the initial period of this work, I was employed at Cambridge University as a research assistant on the VODIS project, funded under the UK Alvey initiative as project MMI/HI 003. My thanks to Frank Fallside and Steve Young of Cambridge University Engineering Department for their support during this phase of the project, and to the other VODIS partners, British Telecom and Logica. I have also profited from discussions with Lee Fedder, Dave Good, George Houghton, Philip Johnson-Laird, Alex Monahan, Caroline Proctor, and Kim Silverman.

The main body of my research, including the results reported in this thesis, have been carried out within the Sundial project. This project is partially funded by the CEC ESPRIT programme, as project 2218. The partners in the project are CAP Gemini Innovation, CNET, CSELT, Daimler-Benz, Erlangen University, Infovox, IRISA, Logica, Politechnico di Torino, SARIN, Siemens and The University of Surrey. My thanks to these organisations for their cooperation, and to the following individuals, for technical cooperation, and for useful if sometimes stormy debates: Eric Bilange, Wieland Eckert, Nigel Gilbert, Marc Guyomard, Paul Heisterkamp, Jean-Yves Magadur, David Sadek, Jacques Siroux, Robin Wooffitt. I have collaborated particularly closely with Jill House, and Scott McGlashan, gaining much in our exchanges.

Thanks to British Airways for permission to use extracts from their recordings.

Thanks to my collegues: Norman Fraser, Simon Thornton, and Trevor Thomas, and others in the Speech Lab at Logica, for providing a stimulating work environment. Especial thanks to Jeremy Peckham, manager of the Logica Speech and Natural Language Group, for appreciating the complementary nature of my project commitments and research goals, and for his generosity when I needed time to catch up with the latter.

Two individuals provided indispensible support with infrastructure. Mervyn Wingfield helped me solve my hardware problems. Roy Patterson made available a room and a MacIntosh, during the critical period of writing up.

I owe a great dept to my external supervisor, Anne Cutler. On innumerable occasions, she has boosted flagging morale, and managed to keep things going, with an impeccable sense of timing. She has been a patient reader of countless drafts, uncompromising in her demands for quality and relevance. Where this work falls short of those requirements, the fault is mine.

Lastly, a big thank-you to Marija, Ben and Jamie, for putting up with so much. This thesis is dedicated to them.

To my family

Chapter 1

Introduction

1.1 Information dialogues

Dialogue is a basic social skill; human infants engage in dialogues of a rudimentary nature before they learn to use language. A competent language user can participate in dialogues, and follow passively the conversations of others.

In the case of human-computer interaction, the purpose of the interaction may be to get the computer to perform an action or series of actions. Of particular interest are *information dialogues*, defined by Bunt (1989) as "dialogue with the sole purpose of transferring (obtaining, providing) certain factual information." Language-based interaction, in particular that where the medium of communication is voice, provides an efficient means of information transfer. Information services may become available to relatively unsophisticated users, provided that commonly used linguistic and dialogic conventions are observed.

1.2 Prosody in dialogue

The distinction between voice and written conversations is not simply a matter of modality. The linguistic register is often different. Writing seems to encourage more formality, more structure, less redundancy. On the other hand discourse markers such as *oh* and *well* belong to the spoken mode. Because dialogue tends to be spoken, and the spoken message is subject to errors and noise in processing, voice dialogues

are more likely to give rise to communication failure and the resulting repair needs. A final difference concerns the prosodic component of a voice message, missing in the written. The role of prosody in information dialogues is the subject of this thesis.

In current 'text-to-speech' systems prosodic decisions are made on the basis of textual information alone. Since a complete linguistic analysis is unavailable, the result is often bland, or inappropriate. The case is different for voice output generators forming part of computational systems that embody linguistic and contextual knowledge. McDonald (1989) points out that a language generation program allows an application to distance itself from language:

> This distance makes it possible to increase dramatically the quality of the texts that are produced by allowing independent influences to enter into the process and their constraints and contributions combined ... A grammatically sophisticated generator will annotate ... conceptual units according to the different ways they could be combined and realized, and will establish a set of decision criteria—rhetorical, semantic and pragmatic—which govern what actually happens.

If we assume that prosodic choices are to a greater or lesser extent affected by internal states, then a computational representation of such states, as they evolve during the course of a dialogue, ought to be capable of informing prosodic choices.

1.3 Aims and scope of this thesis

This thesis sets out to produce a computational account of a speaker in dialogue, with primary attention to how decisions of prosodic import are made, given the speaker's internal representations. The representations may be broadly divided into those which concern the dynamic state of both agents' knowledge about the domain, and those concerning the relationship between the speakers and their expectations of one another. In the domain of prosody, I adopt the commonly-held position whereby information status if indicated is done so by accentual means, whereas pragmatic relations are conveyed using melody. From this vantage point, I shall be concerned with these questions:

1. how salience is conveyed prosodically in dialogue;

2. how dialogue is structured, and what implications this has for the choice of intonation contours;

3. how prosody generation can be integrated into a computational account of language production.

The representations that I use have arisen out of the dual need of answering these questions, and building a working computer model of a language-using agent who performs many other functions which are not necessarily manifested prosodically. Wherever appropriate, I shall attempt to show that decisions with prosodic import are emergent, in the sense that the mechanisms which underlie them must exist on independent grounds.

Language behaviour is an extremely wide field of study; in this work I am primarily concerned with modelling the human agent as producer of language. Although the processes used in production and understanding may be different, it may be that they share, at some stage, common representations—a common (language-dependent) grammar and lexicon—to take the most obvious example. Being concerned with the dynamic behaviour of speakers therefore, I shall have occasion to refer to their activity as listeners. Studies in the field of prosody are also numerous and varied. Prosodic phonology, and its phonetic exponents, have occupied much attention. My work is largely concerned with those pragmatic decisions that are manifested prosodically; representations at a phonological level are therefore only of passing importance. Because however as yet the 'metalanguage of prosody' is a relatively impoverished one about which there continues to be controversy, it is necessary on occasions to argue at a level of phonological explicitness: either in reviewing the evidence, or when discussing the outcomes of choices which affect prosody.

A final restriction on this thesis is that the domain, and thus the sublanguage from which most examples are drawn, is restricted to the application of flight enquiries. This is in part necessitated by the origins of this work within the Sundial project. It has the advantage however of considerably simplifying the technical details of surface-linguistic and semantic representation. Examples cited in the litera-

ture often draw on linguistic material which is in itself problematic for computational linguistics technology. The restriction to information dialogues in a simple domain may prove sufficiently constraining to enable generalisations to be drawn, where the sheer variety of material in less limited domains would diminish this possibility.

1.4 Methodology and notation

1.4.1 Computational modelling

As demonstrated in Chapter 5, the ideas central to this thesis have been elaborated in sufficient detail to form part of a working system for handling human-computer dialogues. This thesis attempts to provide cognitive explanations for observed linguistic phenomena. A computational model can support such an explanation in at least two ways. Firstly, the theory that is being stated can be formulated in precise terms: should any observable consequences arise, these may be tested and used to refute or support the theory. Secondly, a computational model which is implemented constitutes an existence proof that the theory is supported by an effective procedure (eg. Johnson-Laird 1983). In describing such a computational model, however, a trade-off needs to be maintained between achieving sufficient explanatory power, and over-shooting into excessive detail.

Notations and implementation

Woods (1986) points out:

> ...one can axiomatise any rule-governed behaviour that one can rigorously specify, whether its nature is deductively valid or not, logic-like or not. But the difficult issues mostly stem from determining what the behaviour should be and how to bring it about. This work remains whether the medium is predicate calculus or some other representational formalism.

This thesis is an exploration of behaviour, some of which would appear to be rule-governed. I have been eclectic in my choice of notations, and not sought to provide rigorous axiomatisations or proofs. Where some degree of formal precision is required, I use a combination of logic, set theory and the algorithmic notations familiar

in computer science. In Chapter 4 I make use of unification grammar formalisms, as a technique for describing linguistic constraints and their combination, as these operate across levels. A similar notation, based on the notion of typed feature structures, is used for describing conceptual material. In common with other featural notations, these constraint languages allow underspecification. Structures can then be built up incrementally. In Chapter 3 the notation of Communicating Sequential Processes (CSP) is introduced, for describing and reasoning about concurrent processes.

The programs embodying the computational theories described here have been implemented in Quintus Prolog. However despite the apparent declarative nature of Horn-Clause logic, on which Prolog is based, the programming language is far from declarative in its every day use. Prolog execution may be described procedurally, as a process of search through an AND/OR goal tree (Eisenstadt and Brayshaw 1987). Viewed in this way, Prolog is not a particularly appropriate language for describing algorithms, since the state of the search at any stage is not explicitly represented, but hidden for example in the previous choice points for which alternatives remain. I therefore make little use of Prolog in describing algorithms or datastructures.

1.4.2 A notation for describing intonation

The term *prosody* is generally used by linguists to refer to those linguistic features of an utterance which are not segmental (eg. Couper-Kuhlen 1986:3). In the acoustic domain, its parameters are generally taken to be those of fundamental frequency, amplitude, and duration. At an intuitive level, where perceived structure is taken into account, these correspond to melody, stress, and rhythm. However, as Bolinger (1958) pointed out, pitch obtrusion is a more important correlate of stress than increased amplitude. Accepting this view, it follows that the field of intonation, in its narrow sense pertaining to the melody of speech, subsumes within it the study of accent. In this thesis I am concerned with intonation; I ignore that equally important aspect of prosody—metrical structure. The latter is generally taken to be realised in the time domain, and therefore separable from intonation, though of course dependent on it.

English is privileged in the considerable effort that has been devoted over the years to the transcription and classification of intonation. A number of systems have been used, from interlinear notation, in which the relative pitch of each syllable is indicated (Palmer 1924), to *tonetic* (eg. Crystal 1969) and *autosegmental* (eg. Liberman 1978, Pierrehumbert 1980), in which contours are described as minimal sequences of *tonal events*, the latter specifying some aspect of local pitch movement. Although emerging out of different traditions, the British (contour-based) and the American (level-based) approaches have tended to converge on descriptions which are substantially similar. Ladd (1980: ch.1) compares a number of systems from both sides of the Atlantic, and observes a general consensus in approaches, for example in the description of nuclear tones.

In describing and transcribing intonation, I shall follow the practice associated with the tone sequence model, and describe only relevant tonal events. Bolinger (1958) distinguishes between *lexical stress*, the potential of words to have accents on certain syllables, and *pitch accent*, the intonational realisation of accent in an utterance. The tonal events, or *tonal segments* of the autosegmental model correspond in the main to pitch accents; these are supplemented however with segments known as *phrase accents* and *boundary tones*, which are used to define behaviour at the periphery of contours. Tonal segments may be interpreted as targets, possibly associated with local pitch movements. It is possible to derive a pitch contour from these, using a variety of interpolation and smoothing techniques (eg. Pierrehumbert 1980, Silverman 1987). The phonological claim for autosegmental representations is that they define distinctive abstract patterns; the phonetic claim is that these patterns are directly realisable in the acoustic domain.

The advantages of an autosegmental-style representation are not only phonological however. A case can be made for the ultimate equivalence of many if not all tonal sequence notations. However, autosegmental representations are becoming commonly used by a wide variety of linguists, both intonologists (Ladd 1983; Gussenhoven 1984; Hirst 1983) and others whose interest is with the interaction of intonation and other components (Steedman 1990; Bird 1991).

Pierrehumbert's (1980) representations make use of two tonal symbols, **H** (high)

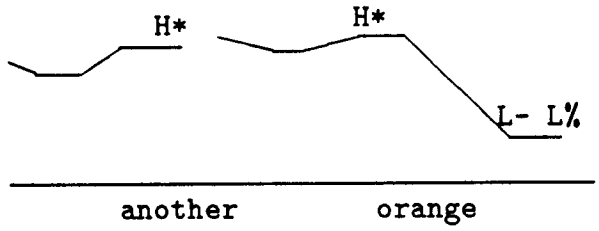
Symbol	Type	Gloss
H	Pitch accent	'high' accent
HL	Pitch accent	(nuclear) fall
$\bar{\mathbf{H}}$	Pitch accent	high with monotone
HH	Pitch accent	high rise
LH	Pitch accent	low rise
HLH	Pitch accent	fall rise
$\uparrow\mathbf{HL}_{\perp}\mathbf{H}$	Pitch accent	deep fall rise
$\downarrow$	Diacritic	downstep
$\uparrow$	Diacritic	upstep
[ ]	Boundary	boundary symbols (default)
$\uparrow$]	Boundary	end boundary: top of speaker's range
$\downarrow$]	Boundary	end boundary: bottom of speaker's range
[$\uparrow$	Boundary	initial boundary: high key
[$\downarrow$	Boundary	initial boundary: low key

Table 1.1: Inventory of tonal symbols

and **L** (low), together with diacritical marks indicating pitch accent, phrase accent, and boundary tone. Certain combinations of tones trigger upstep and downstep, according to rules specific to these phenomena. Ladd (1983) proposes a modification whereby the dichotomous **H/L** distinction is maintained, but phenomena such as downstep and delay are specified as features on accents. I follow the spirit of this proposal, using the symbols '$\uparrow$', '$\downarrow$' to indicate upstep and downstep, and '$\bar{\mathbf{X}}$', where **X** is a tonal accent, to indicate monotone. A complete inventory of the symbols used is given in Table 1.1. The symbols are typed as *pitch accent*, *diacritic*, or *boundary tone*. I follow Ladd in introducing compound symbols such as **HL** for 'fall', where Pierrehumbert would decompose these into pitch accent/phrase accent. Boundary tones are by default not indicated; where an intermediate boundary position needs to be shown, it is generally enough to indicate this using the single symbol ']' (rather than ']['). Marked final boundary tones ('$\uparrow$]' and '$\downarrow$]') are shown when pitch rises to falsetto, or falls to creak. Otherwise boundaries are taken to follow the contour of the nuclear tonal segment. When unmarked, phrase-initial boundaries default to the middle of the speaker's range. Otherwise '[$\downarrow$' and '[$\uparrow$' indicate high and low key respectively. The notation also has some similarities to INTSINT and SAMSINT (Wells *et al.* 1992).

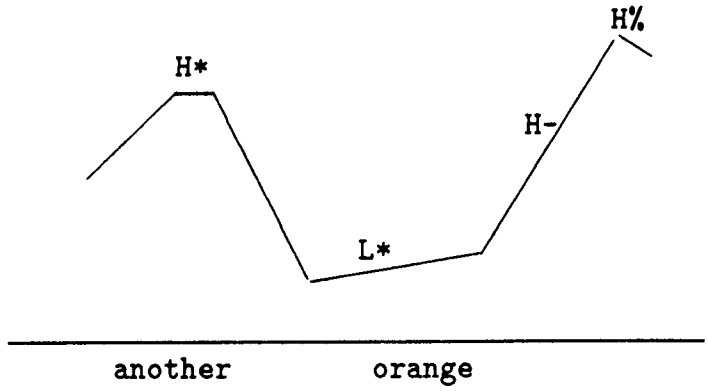
In Examples 1.1–1.3 the (a) versions show stylised copies of contours, marked

using Pierrehumbert's notation (Pierrehumbert 1980: 257, 277, 291); the (b) versions give the same sentences annotated with the notation described in Table 1.1.



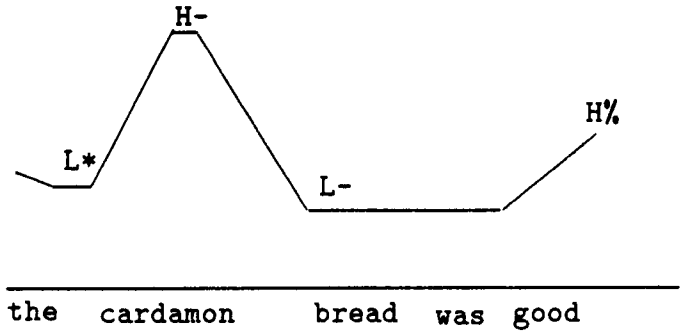
(1.1) a.

b. An**H** other **HL** orange



(1.2) a.

b. An**H** other **LH** orange ↑]



(1.3) a.

b. The **LHLH** cardamon bread was good

I follow Pierrehumbert in marking a 'hat pattern' as **H HL** (cf. Example 1.1). Where the prenuclear accent is at some distance from the nucleus, the intervening syllables 'sag' for **H HL**; I use $\vec{\mathbf{H}}$ **HL** to maintain the level head.[1] Delayed or 'scooped' accents (cf. 1.3) have not appeared in my data; they could possibly be notated more efficiently using a diacritical feature indicating delay (eg. Ladd 1983).

[1] Gussenhoven (1984) introduces a tone-linking rule, whereby a chain of **HL** accents lose their fall, and become high accents with sags in between.

The notation I use for transcription is broad rather than narrow; as a result there may be details which other notations such as that of Crystal (1969) describe better. On one issue at least my usage agrees with that of Crystal and the 'British school': the symbol annotating the nuclear syllable has scope up to the next pitch accent or phrase boundary. This is exemplified in (1.3), where **LHLH** applies not just to *car* but to the remainder of the utterance. In contrast, some autosegmentalists such as Pierrehumbert notate the phrase accent ($\mathbf{L}^-$) and the boundary tone ($\mathbf{H}^\%$) also at the syllable level. My choice is mainly one of convenience; where I depart from it this is because the final boundary tone is at an extreme in the speaker's range.

To facilitate comparison, where several contours are available for a single text, separately, indexed to positions in the string. Position '0' is always the beginning of the string, and '$' the final boundary position. Boundary symbols, unless specifically indicated, follow immediately the word last indexed. So for example, the indexed string:

I'd $_1$like to reserve a $_2$flight to $_3$paris on $_4$monday $_5$evening

together with the contour description:

$_0[^{\uparrow}$ $_1$**H** $_3{}^{\uparrow}$**HL**] $_4$**H** $_5{}^{\downarrow}$**HL**

is equivalent to the utterance:

(1.4) [$^{\uparrow}$I'd **H** like to reserve a flight to $^{\uparrow}$**HL** paris] on **H** monday $^{\downarrow}$**HL** evening

I shall wherever possible use this version of autosegmental notation, even when citing examples which have appeared otherwise notated. Such a straightjacket approach is justified because the phonology of intonation is not one of the main concerns of this thesis.

1.4.3 A notation for describing prosodic focus

It is well known (eg. Brown et al. 1980, Thompson 1980) that prosodic labelling of spoken corpora is difficult, and difficult to get even experienced judges to agree upon. Among other aspects, this applies to the presence and absence of sentence

accents, which are generally taken to indicate prosodic focus. Because of this, when discussing focus I use a notation intended to leave unanswered the question of presence or absence of accents, but which indicates simply that certain phrases are more or less prosodically prominent than others. Thus I use the symbols < ... > round a phrase to indicate relative prominence, and > ... < to indicate relative lack of prominence, with respect to the surroundings. For example:

(1.5) it <seems> to be <ahead of schedule>

(1.6) it ≪is≫ahead of schedule now

(1.7) ...leaves at seven twenty in the evening our time
which is eight twenty >in the evening< their time

The double bracketing in (1.6) is used for especially prominent phrases, which normally would receive some kind of emphatic pitch marking.

The amount of marking used will depend on the degree of detail at which distinctions are needed to be made. Examples 1.5–1.7 are relatively under-marked. A more comprehensive marking might be the following:

(1.5') >it< <seems> >to be< <ahead of schedule>

(1.6') >it< ≪is≫>ahead of schedule< <now>

(1.7') ...leaves at seven twenty in the evening our time
>which is< ≪eight≫<twenty> >in the evening< ≪their≫>time<

Whereas this seems a reasonable extension of the previous versions, it will be observed that any attempt to define a language of text delimiters upon which such a system of marking can be based is likely to fall prey to problems of redundancy and ambiguity. The obvious solution is to impose these bracketings on phrasal structures rather than flat ones. This would have the advantage too of being capable of representing relative prominence at a global as well as at a local level. This approach is the one taken in metrical phonology; I shall resist it, since my principal concern is with not with surface structures, but with the mechanisms underlying focus. The notation as presented, especially when used at a lesser level of detail than in Examples 1.5'–1.7', is sufficient for illustrating phenomena, when default conventions

such as the routine defocussing of closed class words is taken into account. In Sections 4.2.5 and 4.3.3.1 I introduce underlying representations of focal prominence, according to which the output of the focus assignment mechanisms discussed in this chapter may be formatted.

The two notations introduced, pitch-accent-based autosegmental, and bracketed for focus, are complementary. In general I will use the first when contours are being discussed, and the second where issues of focus assignment are at stake.

1.4.4 Corpora

The prosodic data discussed in the literature is largely of the anecdotal kind. Some intonologists maintain their own corpora of examples (eg. Bolinger 1989: 394–397; Couper-Kuhlen 1986). However much of the argument over the interpretation of intonation draws on hypothetically constructed examples and scenarios. There is no doubt some validity in this approach: consider what advances have been made in the study of syntax since Chomsky led the way by advocating the respectability of linguists' intuitions. And as Bolinger warns, unless they are carefully constructed, statistically-based studies may miss those occasions where intonation is being put to special use.

In this work I shall refer where relevant to examples from the literature, anecdotal or otherwise. I shall also on occasions construct my own examples. However, a large part of the data I present is taken from two corpora, which I describe here.

The first, the *Flight Enquiries* corpus, consists of a set of recordings of spontaneous telephone conversations between professional information providers in the employ of an airline, and individual callers. As one conversation (not transcribed) reveals, the employees were aware they were being recorded. The conversations typically are concerned with queries about specific flights; occasionally the queries are less specific requests for information, or requests which fall outside the competence of the information provider, such as enquiries about flight reservations. Selected conversations from these recordings were transcribed by a collegue, Robin Wooffitt, using the notation of Conversational Analysis; indications in the notation of

gaps in the conversation and overlaps are left in where appropriate, but no use is made of them. Where relevant I have transcribed portions of the corpus, using the autosegmental or focus notations described above.

The second corpus—known as the *Swedish Airlines* corpus—arose out of a need to study in more detail the dialogue phenomena as they applied to the even more limited case of information dialogues which could be produced by the computer information service under construction within the Sundial project. The dialogues were designed to provide a variety of linguistic and dialogue-structural forms that appeared to lie within the projected linguistic competence. Phrases and phenomena were based closely on those found in the enquiries dialogues, but in some of the dialogues the task of making reservations was included. An attempt was made to cover the range of dialogue acts that should be within the capability of the agent. Particular attention was paid to repetitions, and the difference between strong and weak confirmations. The texts of the dialogues is given in Appendix A.

Two sets of recordings were made, with two speakers in each taking in turn the part of Caller and Agent.[2] The speakers: MG, JM, JQ and MC were one female and three males. Since the purposes of the recordings were to provide material for the theories described here, rather than to establish empirical results, no effort was made to normalise the conditions of recording or the resulting recordings.

The nine dialogues were transcribed, using the annotated autosegmental notation (see Appendix B). Not all turns were transcribed for every speaker, but an attempt was made to do this in the case of potentially interesting phenomena. Because this work is concerned with formulating a theory of production specifically in the case of the information-provider, somewhat more attention was paid to the utterances of Agent than to those of Caller.

The two corpora are complementary: the first is natural and spontaneous, but contains limited examples if any of a given phenomenon. The second is constrained to be closer to the competence of the linguistic and conceptual mechanisms that I present. It also contains more than one example of phenomena such as repetition,

[2] I thank Jill House and Anne Cutler for organising the two recording sessions.

with the contextual conditions and the speakers varied. Against this must be set the unnaturalness of the task of reading from a script, which several of the readers commented on. Extracts from the corpora which appear in the text are identifiable, in the case of the Flight Enquiries corpus, by a reference identifier from the Wooffitt transcriptions, and in the case of the Swedish Corpus, by cross-identification of dialogue number and move number, using the prefix 'SA'.

1.5 The place of this work in the Sundial project

The work presented for examination in this thesis represents my original contribution to the study of pragmatic and discoursal influences on the production of prosody. Although this work is self-contained, as demonstrated in Chapters 3 and 4, the computer implementation needs to be set within the wider context of the Sundial system, of which it forms part. Sundial is a five-year ESPRIT project, having the aim of providing a spoken conversational interface to information services. The system, in particular the dialogue manager and voice output components, is discussed in Chapter 5. The architecture, notably that of the Dialogue Manager, is the result of joint work. The software modules dealing with dialogue and output generation were implemented by a number of individuals. The Belief Module (Section 5.2) was designed and implemented by me, using Nigel Gilbert's NOOP object oriented package. Included in the Belief Module are those microplanning, or description routines, covered in Section 5.3. I was also responsible for the Linguistic Generator (Section 5.4), and the pragmatic components of prosodic generation (Section 5.5). Together these components cover the principal ideas presented in Chapters 4 and 5. Other components of the dialogue manager, described briefly in Sections 5.1.1–5.1.2, were built by the following individuals: Eric Bilange, Wieland Eckert, Norman Fraser, Nigel Gilbert, Klaus Heussler, Jean-Yves Magadur, Scott McGlashan and Jutta Unglaub. Prosodic realisation has been the work of Jill House.

Discussion of those components for which I am not primarily responsible has been kept to a minimum. The computational model outlined in Chapter 3 presents my reconstruction, in part inspired by the Sundial Dialogue Manager.

The semantic knowledge representation language, SIL (see Section 5.2.1), has been jointly developed with a number of individuals, notably Scott McGlashan, François Andry, and Gerhard Niedermair. Likewise, the lexicon description language, a version of Unification Categorial Grammar (see Section 4.2.2), was developed with the aid of Norman Fraser, Scott McGlashan, François Andry and Simon Thornton. Both of these knowledge representation languages replaced earlier prototypes which I built, the essential difference being one of scale and generality. I have played a consultative role in the development of the new languages, having the principal responsibility for implementing the knowledge bases which use them. The interface language between the Linguistic Generator and the rule-based synthesis component (Section 5.5.3) was the joint work of Jill House and myself.

During work on this thesis, some of my ideas have appeared in joint publications, and in Sundial technical reports. Those publications which directly relate to the work of this thesis are Youd and House (1991) and Youd and McGlashan (1992). The first of these presents in more embryonic form some of the ideas on focus elaborated in Chapter 3. The second paper includes details of the Belief Module description routines and the Linguistic Generator. Those components of both papers which represent my own work, and which are re-presented here, have been extensively revised and elaborated. In addition, my contributions to the following technical reports: Bilange *et al.* (1990), Bilange *et al.* (1991), Youd *et al.* (1990), Youd *et al.* (1991) reference material which is re-presented in this thesis.

1.6 Overview of this thesis

In Chapter 2 I review past research in the fields of dialogue modelling, discourse modelling, language production and intonational meaning. While a wealth of analyses of prosodic phenomena are available, little work has been done so far to tie them into a computational model of production.

Chapter 3 examines the Flight Enquiries corpus, with a view to determining the kinds of phenomena that a computational model of dialogue should account for. Such a model is then elaborated, in terms of an architecture of communicating

agents. This leads to an abstract featural characterisation of moves in a dialogue, as 'dialogue acts'. Prosodic transcriptions from the Swedish Corpus are then examined, with a view to establishing whether intonational contour can be said to depend on dialogue act features. It proves difficult to account for the variability in intonation contours in terms of the model; this suggests that factors not modelled, such as attitude, may be involved.

In Chapter 4 I present a computational account of semantic focus, as it underlies prosodic accent. I do this at three levels: surface-structural, conceptual, and intentional. There is some evidence that speakers re-use a certain amount of previous material, including lexis and surface structure. I propose an account of how this is done, and show that it can be of benefit to both speaker and listener. At the conceptual level, reuse of accessible material is often accompanied by relative loss of prosodic prominence. I elaborate a model of accessibility to account for prominence patterns. This is extended to cover cases of contrastive prominence, corresponding to disparate belief states. Finally, I investigate how a speaker's intention to mark certain parts of the utterance may be associated with prosodic prominence and local contours.

Chapter 5 describes and illustrates the implementation of the focus assignment components, as part of the description and generation routines of the Sundial system. I show first how the architecture of the Sundial dialogue manager corresponds to that specified in Chapter 3. I then describe in detail the implementation of the discourse-modelling component, and of the message output components, and the working of prominence assignment. This is illustrated with a detailed example.

Finally in Chapter 6 I discuss some of the implications and extensions arising out of this work.

Chapter 2

Review of previous work

2.1 Introduction

Four areas are relevant to this work. Firstly, theoretical and computational models of human dialogue place the utterances produced by speakers within a framework of interaction, in which common conventions allow speakers and listeners to communicate goals or mental states. Secondly, theories of internal representation of information, and their accessibility to speakers during discourse are required in order to account for the context-dependence of language. Thirdly, cognitive and computational accounts of language production itself examine the constraints on the speaker, and how these can be modelled. Lastly, it is necessary to examine how these contexts: dialogic, discoursal and linguistic, affect prosodic decisions.

2.2 The nature of dialogue

2.2.1 Theoretical and empirical approaches to dialogue

Dialogue is a structured social event in which individuals interact, and as a result of interaction, may change their internal representations. In pursuing the pragmatic functions of prosody, two questions must be asked:

1. What is the nature of dialogue structure, and how is it apparent, either to the speakers during the course of the dialogue, or to the analyst?

2. What factors influence a speaker's behaviour in dialogue, from one moment to the next?

The questions are clearly interrelated: a conversational agent needs to maintain some internal representation of the current context, so that he can deal in a rational, coherent manner with conversational events produced by himself or by the interlocutor. Conversely, any structure that can be shown to exist, it may be argued, exists on functional grounds, namely those of enabling speakers to so orient themselves. In this section I investigate theoretical and empirical approaches to these questions. Approaches based on speech-act theory have to a large degree been theoretical. They have attempted to characterise the relation between speech events, and change in speakers' internal representation. Approaches in the tradition of conversational analysis, on the other hand, have tended to remain pretheoretical in their choice of units of analysis. They have put the emphasis on observation, followed by functional accounts closely tied to the data.

Speech act theory is founded on the assumption that we 'do things with words' (Austin 1962). In its most explicit formulation (eg. Searle 1969) the effect utterances are intended to achieve on their hearer (their *illocutionary force*) is linked to necessary and sufficient conditions for their performance. For example, a question has the *preparatory conditions* that the speaker does not already know the answer, and that the hearer is not expected to provide the answer without being asked, and the *sincerity condition* that the speaker should want to know the information. A final *essential condition* defines what a speech act counts as—in this case an attempt to elicit the relevant information from the interlocutor. Speech acts are therefore intimately bound up with goals and beliefs, assuming as they do goals and beliefs on the part of the speaker, and having at least one effect on the beliefs of the hearer, namely that those goals and beliefs conventionally associated with the use of that speech act are held by the speaker. Searle, and a number of others (eg. Bach and Harnish 1979) have attempted to classify speech acts, using categories such as:

(2.1) **directive:** an attempt by the speaker to get the addressee to do something

Speech Act theory has commanded considerable attention among linguists and computational linguists. However, some theoretical difficulties exist (eg. Levinson 1983: 241ff). Particularly pertinent to the application of the theory in computational dialogue systems, is the question of how it accounts for the relation between surface structures and illocutionary force. According to the *literal force hypothesis* (Gazdar 1981) every speech act has an illocutionary force which is directly dependent on surface syntactic features such as mood, or on the presence of some 'performative' component. Such a version of the theory would have it that an utterance such as *can you open the door* is literally marked as a question, and only via some process of pragmatic inference is the intended force, or *indirect speech act*—in this case a request—derived. This lays a considerable burden on the interpretative process, especially since it is apparent that only a minority of speech acts that occur in natural conversation are direct according to the definition.

In its precision, at least so far as the conditions for successful performance of an act are concerned, Speech Act theory lends itself well to attempts at formalisation.

Cohen (1978) applied AI planning formalisms to the description and implementation of speech acts in computer programs. He proposed that conversation be viewed as a sequence of actions, in which the participants affect each others' beliefs and goals, via speech. Such a sequence could be extended to include nonlinguistic actions aimed at fulfilling an agent's intentions, such as getting a door opened. Within the planning paradigm (eg Fikes and Nilsson 1971) plans had been modelled as sequences of operators generated by a problem solver in order to achieve some specified change in the state of the world. Cohen observed that speech acts could be considered among such operators, since they may be used to affect the beliefs and intentions of agents in cooperative planning. For example, a request is defined as an operator with preconditions:

(2.2)
AGENT believe (*RECIPIENT can do ACT*)
AGENT believe (*RECIPIENT believe* (*RECIPIENT can do ACT*))

—ie, the speaker believes the hearer can perform the request, and is aware of this ability; and effects:

(2.3) *RECIPIENT believe (AGENT believe (AGENT want ACT))*

—which embodies the condition of hearer uptake of the speaker's intentions in performing the act. Cohen remarks that non-obviousness (ie, that the hearer would not be expected to perform the act as a matter of course) and sincerity (that the speaker intends the hearer to perform the act) are both integrated into the planning process. Important features of the approach pioneered by Cohen are the formalism which makes use of embeddable attitude operators such as *believe* and *want*, and the requirement that the effects of speech acts depend on their successful uptake. In fact for their intended (or *perlocutionary*) effect on the interlocutor to take place, further conditions on his willingness and ability to act are needed.

Bunt (1989) extends Cohen's approach. Bunt observes that the illocutionary and perlocutionary effects of communicative acts, which he terms *dialogue acts*, are derivable from their appropriateness conditions. For example

> S wants to know whether some proposition p is true
> S suspects that H knows whether p is true

are basic prerequisites of asking a question. The notion of appropriateness condition is used to define a hierarchy of dialogue acts; these fall into three basic types: *questions, answers* and *informs*. Figure 2.1 shows part of the hierarchy for questions. Dialogue act types are shown in capitals; sets of appropriateness conditions are represented by sentences. Appropriateness conditions are inherited down the tree. Bunt handles lack of certainty about the interlocutor's beliefs, using the operator *suspect* in the place of *know*. In the case where, for example, an agent wishes to obtain confirmation for something he believes to be true for the interlocutor, he may use the POSI-CHECK dialogue act. In Section 2.2.2 I consider in more detail some of the work of the 'planning school' of dialogue modelling.

The conversational analysis (CA) tradition has been concerned with the detailed analysis of naturally-occurring conversations, in recorded or transcribed form. In

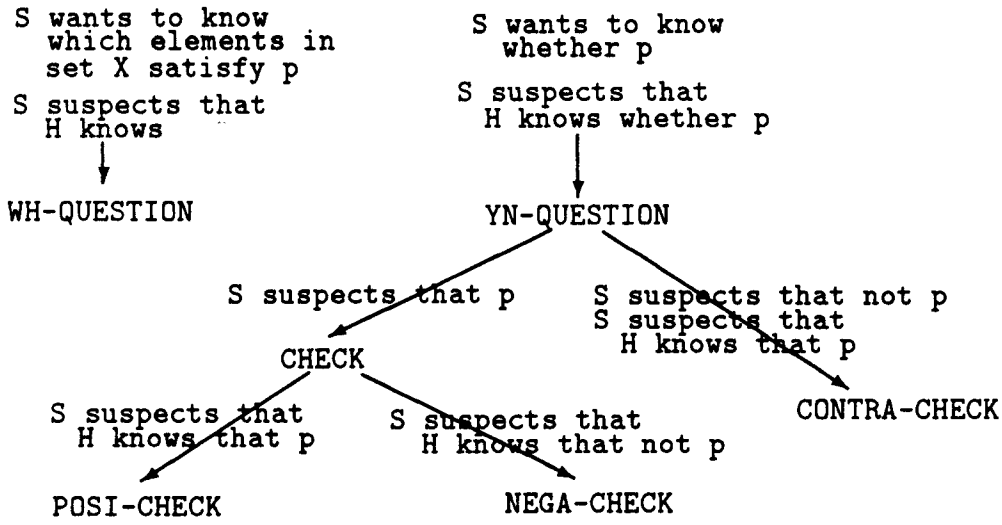


Figure 2.1: Dialogue acts characterised by appropriateness conditions

isolating items for scrutiny, attention has been placed on those phenomena which the conversational participants themselves orient to, as well as to those which can be shown to directly relate to those issues faced by agents engaged in online conversation. One major result concerns the organisation of turn-taking. What are the mechanisms (or rules) whereby agents collaborate to ensure that speaker transition takes place smoothly? Sacks, Schegloff and Jefferson (1974) suggest that these depend on speakers organising their utterances into *turn-constructional units* and listeners being capable of deciding when a unit is complete. The rules then determine who will hold the floor after the current unit: an interlocutor, if he is somehow selected by the speaker, or if he self-selects; failing that, the speaker himself may continue. The rules predict for example that overlaps between speakers may occur if several compete for self-selection, or if the location of the end of the speaker's turn-constructional unit is ambiguous.

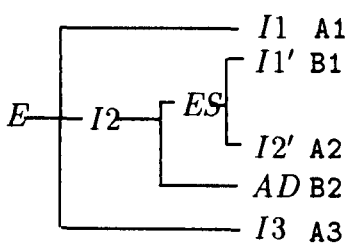
Turn-taking is an example of the *local organisation* of conversation. Equally important for our purposes is the phenomenon of the *adjacency pair.* These are two-person exchanges in which the second part in some way relates to the first part, as for example questions/answers, summons/responses, requests/compliances. However like other rules in CA, the rules defining what count as adjacency pairs are normative rather than prescriptive. That is, failure to produce a second part of

the appropriate type is not an error, though it may lead to pragmatic inference, as is the case when an interlocutor fails to answer a question. It is in fact possible to associate with many patterns of organisation a dispreferred as well as a preferred outcome. Thus a dispreferred outcome in the case of a question would be a 'don't know' answer; in the case of a request, inability or refusal to comply. Levinson (1983) suggests that the indirect nature of many requests such as *do you have hot chocolate* may have arisen because lack of compliance with a direct request is socially dispreferred.

While speech-act based approaches have dealt with utterances in isolation, at the level of agents' beliefs and intentions, conversation analysis has placed much emphasis on observable, structural phenomena. A number of hybrid approaches have attempted to integrate speech acts as units within larger structures. Sinclair and Coulthard (1976), and Moeschler (1986,1989), have the exchange as a unit. Moeschler also introduces rhetorical relations between constituents, which can be marked as dominating or subordinate. Consider the following example (translated from Moeschler 1986):

(2.4)
A1 Have you got the time?
B1 Haven't you a watch?
A2 It's stopped.
B2 It's eight o'clock.
A3 Thanks.

The units of structure are the *exchange*, the *intervention*, and the *speech act*. Exchanges consist of interventions, which may recursively contain exchanges. Moeschler provides the following analysis of (2.4):



The outer exchange, E, is partitioned into an initiating question ($I1$), a response ($I2$) and an evaluation ($I3$). $I2$ in turn is complex, comprising a subordinate exchange (ES) and a dominating act (AD). Embedded exchanges are thus handled in the

grammar. A similar approach is taken by Wachtel 1986; units and rewrite rules governing them are shown in (2.5):

(2.5)
$$\begin{aligned} \textsc{conversation} &\rightarrow \textsc{dialogue}\ [\textsc{dialogue}]^* \\ \textsc{dialogue} &\rightarrow \textsc{exchange}\ [\textsc{exchange}]^* \\ \textsc{exchange} &\rightarrow \textsc{act}\ [\textsc{dialogue}]\ \textsc{act}\ [\textsc{dialogue}] \end{aligned}$$

Wachtel also makes use of features on nodes; these percolate and trickle through the dialogue tree according to various inheritance principles: for example, the (discourse) topic is defined at the dialogue level, and inherited by exchanges and their immediately descendant acts, but not by embedded dialogues.

Houghton (Houghton 1986) combines an intentional model comparable to Cohen's with a syntactic treatment which allows sequences to be treated as structural units. He does this by defining preconditions and effects, not on individual acts, but on sequences of utterances known as *interaction frames.* To accomplish some change in goals and beliefs interactively, an agent must initiate such a frame. While the frame is still active, the agent and his interlocutor take up the initiatory and reactive roles, as defined for that type of sequence. Perlocutionary effects are modelled by the successful culmination of the interaction frame. For example, the interaction frame for MAKE_KNOWN is defined to have preconditions as in (2.6) and effects as in (2.7).

(2.6) $know(initiator, not(know(addressee, proposition)))$

(2.7) $know(addressee, know(initiator, proposition))$

Once this has been chosen, on the basis of the initiator's intentions and beliefs, the addressee will if cooperative become the respondent, and both will act and update their representations according to the instructions laid down for their respective roles.

2.2.2 Plan-based accounts of dialogue

Cohen's OSCAR (Cohen 1978) provided a computational model of his formalisation of speech acts, as plan operators. Where its plan for action required assistance from an interlocutor, OSCAR produced utterances corresponding to the required speech acts. In a complementary approach to that of Cohen, Allen (Allen and Perrault 1980, Allen 1983) produced a computer program in which the system was concerned with modelling the user's underlying plans and goals, using the evidence of speech acts recognised. Allen used a *plan inference* system to provide a rational account of some phenomena in cooperative dialogue, as in the following exchange:

(2.8) patron: when does the Windsor train leave?
clerk: 3:15, at gate 7.

Here the clerk provides, gratuitously, unsought-for information. Allen suggested this was because s/he is able to infer that the patron has the plan of catching the train. Reconstructing the plan, the clerk may detect that the patron needs to leave from a certain gate, and may not know which. The additional information is designed to circumvent this potential obstacle.

Litman (Litman and Allen 1985) considered the case of *clarification subdialogues* as a special case of obstacle detection. If an agent is trying to reconstruct his interlocutor's plan, and finds some information missing, he may plan for a clarification question. In order for her system to reason about plans, Litman included the latter as objects in her ontology, and allowed for *metaplans*, such as SEEK-ID-PARAMETER, which came into action when a parameter of some plan was not known. Instances of plans were kept on a stack, so that when the purposes of a plan were achieved, it could be popped from the stack. The input speech act could be used to continue the current plan inference, in one of three ways, either by continuing the top (suspended) plan on the stack, or introducing a clarification metaplan for some plan on the stack, or by constructing a new plan. Like Allen, Litman used domain-specific expectations of speakers' goals.

A further application of plan recognition has been in the interpretation of sentence fragments. Carberry (1985, 1988) remarked that a range of pragmatic infer-

ences may be associated with these; compare:

(2.9)
A : The Korean jet shot down by the Soviets was a spy plane
B : With 269 people on board?
B': With infrared cameras on board?

While utterance B' is concerned with obtaining (or confirming) additional information, B expresses doubt about the underlying proposition. Carberry poses the question: what knowledge does an information provider need, to cope with such fragmentary utterances? She identifies four major components: (i) a task-related plan, sufficiently explicit to account for the information seeker's goals as revealed in the previous discourse; (ii) a representation of shared beliefs; (iii) a set of anticipated discourse goals, based on general conversational principles; and (iv) processing knowledge, including plan-recognition strategies and focussing mechanisms. Processing fragments takes place as follows. A discourse component suggests discourse goals that the information seeker *IS* might be pursuing. These are stacked, and serve to schedule responses to *IS*'s utterances in such a way that conversational principles are conformed with. For example, having asked a question, the information provider may push onto the stack the discourse expectation ANSWER-QUESTION. Carberry provides rules relating each discourse expectation to a family of possible discourse goals, which *IS* might use to satisfy that expectation. Plan recognition techniques, similar in essence to those of Allen, are used to propose possible attachment points for IS's incoming utterances, to a *context model* representing that portion of IS's plan derived so far. Conflicts between alternative choices are resolved using the possible discourse goals, together with a focussing mechanism that attempts to enforce local coherence by preferring associations in or close to the area identified in the preceding discourse as the current focus of attention.

In the plan-based models cited so far, a single agent was modelled. Interaction was presumed to be with a human user. By contrast the implementation described by Power (1979) is symmetrical; both agents in a dialogue are represented by independent communicating computational processes. Two 'robots' are initially equipped with possibly disparate sets of belief and goals, in a 'world' consisting of the two robots themselves, a door and a bolt, and simple 'laws of nature' governing

action in that world. The robots, two identical programs set up with different initial conditions, have as resources the ability to plan privately, and the ability to initiate and participate in *conversational procedures.* These are used as a basis for conversational structure, and may themselves call embedded conversational procedures. An example:

(2.10) CONVERSATIONAL PROCEDURE ASK(S1, S2, Q)

1: S1 composes a sentence U which expresses Q as a question, and utters it.

2: S2 reads U and obtains a value for Q. He records that S1 cannot see the object mentioned in Q, and then inspects his world model to see if Q is true. If he finds no information there he says I DON'T KNOW, otherwise YES or NO as appropriate.

3: S1 reads S2's reply. If it is YES or NO he updates his world model appropriately. If it is I DON'T KNOW he records that S2 cannot see the object mentioned in Q.

Both speakers share the same inventory of conversational procedures, and each of these is intended to achieve a given purpose for the initiator. For example the conversational procedure ASK is initiated by a speaker whose purpose is to obtain information. Then in order to get the addressee to know which conversational procedure he is expected to take a role in, the initiator has to announce it, as for example, for ASK:

(2.11) May I ask you a question?

In the initial stage of a conversation, the robots, being cooperative, agree to pursue a common goal, such as that of robot1 being on the same side of the door as robot2. The robots develop in parallel identical plans, planning in private, but identifying where necessary those parts of the planning tree which can only be accomplished cooperatively. For these, conversational procedures are used, both to delegate responsibility and to announce its results. For example, to achieve the goal of having the door open, one agent may be assigned the responsibility of pushing it, and later inform the other agent of the new state of affairs. Allowing planning and the execution of plans to interact, as well as the modelling of two agents, marks Power's work as different from the plan-recognition work considered so far.

The implementation of Houghton and Isard (Houghton and Isard 1986, Houghton 1986) is close to that of Power, with interaction frames (discussed in the previous section) replacing conversational procedures. As with Power's program, both agents are modelled and there is therefore no interaction with a human user. The system of Smith *et al.* (1992) also takes the approach that dialogue emerges out of the problem-solving process: interaction takes place when an agent's knowledge is insufficient. The system, which includes speech input and output, is capable of allowing both system and user initiatives.

How suitable is the planning approach as a model of dialogue? In terms of explanatory power, it might be preferred over more syntactic approaches such as that of Moeschler. The model concentrates on the goals of agents and their relations in plans; structure such as embeddedness comes out as an emergent property, rather than needing to be defined in some grammar. Grosz and Sidner (1986) propose an architectural account which brings together surface linguistic structure and underlying *intentional structure*, and treats the former as a derivative of the latter. However Suchman (1987) takes issue with the view that plans form a suitable model underlying action in dialogue. She focusses on the relation of plans to their execution, and proposes that the performance of actions is better explained in terms of local interactions between an actor and his environment; actions so described she refers to as 'situated actions'. According to this view, plans are powerful rationalisations of courses of action, which may be used as a resource before action, in much the way that we look at a map before starting a journey.

2.2.3 Summary

Conversing is the archetypal form of speaking (Levelt 1989: 64). To account for this behaviour requires more than any attempted enumeration of possible conversational sequences can provide. In common with the analysis of utterance tokens, analysis of conversation needs to be based on underlying structure. From a logician's viewpoint, and also computationally, a sensible approach to this problem is to characterise the internal states of agents, then to describe how these states become modified as

the result of events during the course of conversation. Searle's characterisation of speech acts in terms of felicity conditions for their successful performance, and the formalisation of this by Cohen and followers have led to a number of computational models. These take as their starting point the beliefs and goals of agents, and define speech acts as operators on these beliefs and goals. Planning—the derivation of sequences of actions to achieve desired effects—is seen as the motivating force both in the production of speech acts, and in their interpretation. In the latter case, it is the interlocutor's plan that needs to be inferred. A number of dialogue phenomena, from the embedding of clarification sequences, to the interpretation of elliptical or phrasal expressions, fall out well from these accounts.

The range of dialogue phenomena is however somewhat limited; actions pertaining to repair, or acknowledgement of responses, are not handled easily within the speech-act-as-operator model, though Bunt has succeeded in extending it to a range of dialogue phenomena, such as confirmations and corrections. However the dynamic nature of dialogue, as an unfolding of sequences of particular kinds, is not explicitly represented. By contrast, approaches which make use of the structural nature of dialogue are capable of accounting for moves such as backchannel utterances for which it is less easy to formulate an intentional basis. They also offer explicitness about dialogue structure, and may entertain the idea of structure above the level of the exchange. As Wachtel suggests:

> "By taking into account configurations of features at nodes, one can isolate ... such elements as 'the last but one topic discussed by the previous user' or 'the first point in this conversation that needed clarification' (Wachtel 1986:39)

It should be pointed out however that meta-references of this kind are extremely rare in spontaneous conversation.

A third approach, that of Conversational Analysis, has sought to emphasise the creativity of an agent dealing with on-line communicative situations. Analysis at a considerable level of detail has provided insights into phenomena neglected by other accounts, such as openings and closings. Conversational Analysis is wary of rules, and at pains to emphasise their normative nature. Nevertheless rules and

terminology have found their way into computational accounts. Cawsey (1991) describes a system in which discrepancies of belief between agents are handled by a reason maintenance system, while 'repair' messages concerning such discrepancies are scheduled according to the normative sequencing properties proposed by Schegloff et al. 1974. Hirst (1991) in fact suggests that so long as rules formulated by Conversational Analysts are declarative, and thus capable of being manipulated by agents at a meta-level, computational treatments are possible. The case for this however is far from proven.

Many existing dialogue systems inspired by Cohen's formalisation of Speech Act theory have been successful in providing coherent theories of agents as both language understanders and producers. There has however been a bias towards understanding, and towards issues of pragmatic inference which are of theoretical interest. The TENDUM system (Bunt *et al.* 1985), which focusses on information dialogues, negotiation and repair, and is partially empirically based, possesses many of the features of the dialogue system that I explore in Chapter 3.

All systems that I have discussed, including those in which dialogue structure is explicitly represented, have in common a division between the representation of information, and the operations or messages which result in the manipulation of that information. Those accounts that are prepared to define units agree on the speech act/speech event/communicative act/dialogue act as the basic unit of representation. Where they differ is in the functional load that they put on this unit. Dialogue grammars, or the interaction frames of Houghton, may represent dialogue events which are insignificant in terms of their effects on goals and beliefs, but which serve as slot fillers in a given sequence.

However, little has been achieved in the elucidation of the relation between surface sentences and phrases, and these acts. The problem is partly one of interpretation; as we have seen, the communicative function of many sentences is conventional, or needs to be arrived at via pragmatic inference. This need not be a problem for accounts of the relationship between dialogue acts and intonational contours, since it might be argued (and has been) that the latter are more directly indicative of illocutionary force. Mismatch of units, however, remains a problem both for accounts

of sentential content, and accounts of intonation. This is because turns, linguistic units and dialogue acts need not be in one-to-one correspondence.

One other factor that has received relatively little attention is the perception by a speaker of his relationship to the interlocutor. Levelt (1989) cites experiments by Herrmann (1983) in which subjects requested objects of one another. It was found that the amount of deference (linguistically) in a request was directly related to the speaker's view of his own legitimacy in making that request. Similarly, in the classroom dialogues investigated by Sinclair and Coulthard (1975) it was found, not surprisingly, that a basic asymmetry existed in the structural positions occupied by utterances of teachers and pupils.

2.3 Discourse coherence

Communicating agents need to keep their own internal record of the discourse thus far; otherwise every utterance must introduce everything anew. So much is implicit in the accounts of dialogue structure reviewed: for example, an utterance counts as a response in so far as the speaker can be shown to be reacting to some previous initiative. But the contextual representations that agents hold in common must also contain a record of what has been said. This might be at the surface-linguistic level: a simple log of every utterance. Or it could be at a deeper level, in which information is stored as tokens in some internal model of the world, as affected by the discourse. There is evidence from studies of human memory, and behaviour in discourse, that the deeper form of representation is the one which persists, and which agents use as a resource during conversation. In Section 2.3.1 I review the evidence for discourse models and discourse model accessibility, and discuss theories of how the discourse model is represented. In addition to their use of long-term discourse memory, it has been shown that over a relatively short span, speakers and listeners make use of surface linguistic information. The evidence for this is reviewed Section 2.3.2.

2.3.1 Discourse model accessibility

Language processing agents build internal models which go beyond what was actually said. We are at pains to make sense of the world as presented to us through language; we will, for example, readily accept the description "the waiter" if the current discourse context concerns a restaurant, even if that individual is being mentioned for the first time. These model-building abilities were dramatically illustrated in an experiment by Bransford, Barclay and Franks (1972). Subjects presented with the sentence:

(2.12) Three turtles rested on a floating log and a fish swam beneath them.

readily make the spatial inference that the fish swam beneath the log. In a similar vein, it has been shown that memory for gist outlives memory for the exact verbal expressions used. Discourse models "are either part of or else intimately linked to broader models of the world" (Hankamer and Sag 1984). We may therefore conceive of them as populated by tokens, or *discourse entities*, which correspond in some way to the entities, real or psychological, that they represent. Additional evidence that humans model the world of discourse in this way comes from the use of pronomial anaphora and referring expressions (eg. Halliday and Hasan 1975; Hankamer and Sag *op.cit.*). These can often be interpreted as referring to previously mentioned discourse entities; moreover, the antecedents are not necessarily noun-phrases, but may be complete sentences denoting events:

(2.13) The children asked to be squirted with the hose, so we did *it.* (Hankamer and Sag 1984:327)

As further evidence that our discourse models correspond to internal representations of the world, Hankamer and Sag point out the interchangeability of anaphoric expressions, between discourse and deictic contexts.

Investigations of the make-up of discourse memory have been motivated largely by the need to provide a unified account of patterns of coherence over stretches of discourse, such as pronominalisation of recent entities, and deaccenting of 'given' entities. The terms 'given' and 'new' were introduced by Halliday (1967), in order

to account for the distribution of intonational prominence within utterances. New information is that "not ... recoverable from the preceding discourse". Halliday and Hasan (1976) extend the notion of given to include situationally recoverable material. Chafe (1974) combines Halliday's idea of recoverability, with the notion of assumed presence in the listener's *consciousness*, which he takes to be of limited capacity, so that discourse entities lose their 'given' status if not mentioned for some time, or if the scene or situation changes. Experiments suggest that this is indeed the case. Clark and Sengul (1979) found that readers interpreted an anaphoric expression more rapidly if the antecedent was in the previous clause, as opposed to two or three clauses away.[1]

Further empirical work on discourse accessibility concerns the status of discourse entities which are referred to anaphorically, although not previously mentioned. Clark and Haviland (1977) considered the case of *bridging inferences*; these are supposed to occur when a definite referring expression is used with no clear antecedent in the preceding discourse, as in (2.14):

(2.14) a. Mary unpacked the picnic things
b. The beer was warm

The phrase *the beer* is not mentioned in (a), but can be felicitously referred to in (b). It is an example of an inferrable discourse entity, according to the taxonomy of Prince (1981). Clark and Haviland conjectured that readers would take longer to process cases such as (2.14), where inference was required, than those such as (2.15) where an antecedent had been explicitly introduced into the discourse:

(2.15) a. Mary unpacked the beer
b. The beer was warm

Experiments, in which subjects were timed reading the second sentences of pairs such as (2.14) and (2.15) showed conclusively that this was the case.

Sanford and Garrod (1981) conjecture that such inferences may be confined to the discourse topic, or *scenario*, under consideration. Scenarios have typical discourse

[1]They made the additional finding that there was a marked discontinuity in retrieval times between the recent and non-recent cases. This provides further evidence for the case put in Section 2.3.2, that (short-term) surface memory and (longer-term) discourse memory are organised differently.

entities associated with them, such as the waiter in a restaurant, or the conjurer at a birthday party. Experiments (again, reading) show that once the limits of a scenario (in time, or in space) have been passed, subjects show greater difficulty in referring to scenario-dependent entities.

Researchers building computational models of memory, especially those concerned with natural language, have used database and knowledge representation techniques to implement discourse models. This was done for example by Winograd (1972), who built a program which conversed about a world inhabited by three dimensional geometrical shapes. The current state of the world was displayed visually; the discourse model at any stage consisted simply of its internal representation. A more flexible approach to representing Natural Language semantics, has been the use of *semantic networks.* Bobrow and Webber (1980) for example use the KL-ONE language (Brachman and Schmolze 1989) for modelling natural language in dialogue. This representation, like many in AI, is semantic-network based, uses inheritance, and is capable of distinguishing objects and classes at various levels of specificity. Linguistic events are then interpreted as operators which affect this structured model.

Such computational models of memory, together with the *script*-based representations of Schank and Abelson (1977) are capable of making bridging inferences. Less straightforward is the modelling of accessibility. Grosz (1977, 1981) analysed task-oriented expert-apprentice dialogues, and found that the set of possible antecedents to a referring expression such as "the screw" depended on the current subtask focussed on in the dialogue. This meant that if a task was returned to after an embedded subtask had been dealt with, a referring expression could refer back to an earlier, not recently-mentioned antecedent. Grosz's (semantic-network-based) model partitioned the discourse model into *focus spaces*, according to the accessibility of discourse entities during different phases of the task. The current focus space was termed 'active'; those above it in the task hierarchy, 'open' focus spaces. Interpretation of a referring expression was handled by directing search for an antecedent firstly in the active and open spaces.

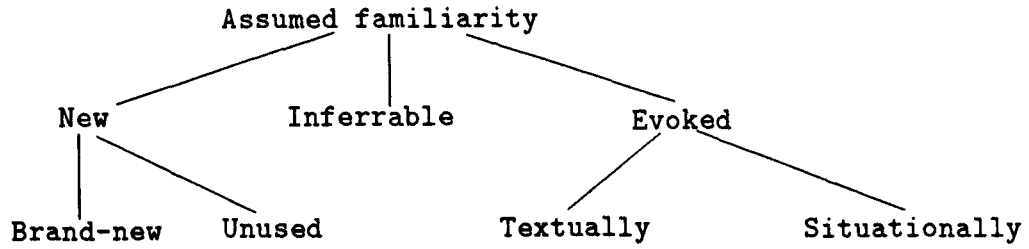


Figure 2.2: Prince's taxonomy of given and new discourse entities

A number of attempts have been made to present unified theories of discourse model accessibility. Prince (1981) attempts to clarify the notion of 'Givenness'. This is treated by Halliday in terms of assumed *recoverability* by the listener; Chafe uses the more limited notion of 'in the listener's consciousness'; if we add the sense of 'shared knowledge', used for example by Clark and Marshall (1981), this amounts to a very imprecise use of the term 'given'. Prince seeks to remedy this by classifying referring expressions according to the way in which they purport to address or update entities in the (assumed) hearer's model. Her classification is shown, in simplified form, in Figure 2.2. A discourse entity may be entirely new to the addressee, or it may be dormant, not having been addressed during the current discourse. An entity which has been evoked, on the other hand, may have been explicitly referenced in the discourse; on the other hand, it may form a part of the discourse situation in which speaker and hearer find themselves. Between new and evoked entities, are those inferrable (or scenario-dependent) entities whose introduction can almost be taken for granted, given the current discourse model.

A slightly different perspective is taken by Levelt (1989), who classifies an agent's model into (possibly intersecting) subsets, according to the source of information: *common ground* contains the information which a speaker can assume his interlocutor to know about as a matter of course, for example because of a shared social situation; the *speaker's contribution* and the *interlocutor's contribution* contain that information which the speaker and interlocutor have introduced, respectively. Finally, there is information which the speaker may potentially introduce at some time in the future. Levelt defines the discourse model proper to consist of the sum of the

speaker's and interlocutor's contributions.

Ariel (1989) proposes that linguistic resources are systematically used to mark discourse entities for accessibility. This is in line with Hankamer and Sag's (1984) work on deep and surface anaphora. According to Ariel, the linguistic encoding of accessibility ranges from high-accessibility markers, examples of which in English are unstressed pronouns, and zero anaphors, through medium-accessibility markers, such as stressed pronouns, and demonstratives, to low-accessibility markers: typically proper names and definite descriptions. Accessibility may depend on a number of factors; among these, Ariel cites: (i) the *distance* between antecedent and referring expression; (ii) *competition*—the number of competitors having the same role as the antecedent; (iii) the *saliency* of the antecedent; for example, whether or not it is the current topic; (iv) *unity*—whether the antecedent belongs to the current world/frame of reference, or in textual terms, to the same paragraph or discourse segment. Drawing on cross-linguistic evidence, Ariel surmises that this accessibility scale is universal, but in part conventionalised with respect to the language in question. As evidence of a universal tendency, it can be shown that whereas high accessibility is marked by using attenuated forms low on informativity, less accessible entities are expressed using more informative forms. Similar ideas are explored by Givón (1992), who argues that such linguistic devices are used in language production to convey not only referential accessibility, but also *thematic importance*, in terms of the activation of material with cataphoric potential. Expressions such as nominals may be interpreted as instructions for cognitive operations on discourse referents, topical or otherwise. Zero-anaphors or pronouns mark continuation of the currently active "file", or topic, according to the iconicity principle:

> Information that is already activated requires the smallest amount of code. (Givón 1992: 25)

2.3.2 Accessing and reusing surface forms

A number of researchers have found that, whereas listeners/readers may have a good recall for gist, memory for surface forms is short-lived. Levelt and Kelter (1983), for example, recorded conversations with shopkeepers, who were systematically asked

one of:

(2.16) a. What time do you close?
b. At what time do you close?

Although there is no difference semantically between the two sentences, the form of the answers was found to correlate significantly with the question forms. However, if other material such as an explanation was interposed between the question and response, the correlation dropped to chance level. Levelt and Kelter concluded that while surface recall had an effect on speakers' productions, its effect was limited to adjacent clauses.

One hypothesis concerning the value of recent utterances, to speakers or listeners, is that there may be a reduction in processing load, if the results of previous processing continue to be available. Frazier *et al.* (1984) studied how parallelism affected readers' performance. Subjects were presented with sentences formed out of conjoined clauses, for example:

(2.17) John telephoned the library and his friend telephoned the doctor

There was strong evidence that subjects processed second clauses which exhibited a high degree of parallelism with their first parts, faster than comparable material, and with improved comprehension. The effect in fact carried over to cases where the parallelism was semantic: Example 2.17 would take longer to process than a similar sentence in which both direct objects were of the same animacy. Frazier et al suggested that these results could be partly accounted for by the assumption that intermediate representations were primed; to the extent that processing a subsequent segment could take advantage of existing representations, the cognitive load would be diminished.

In a similar vein, Bock (1986) studied the tendency of speakers to produce forms that were congruent with just-uttered sentences. Subjects were primed by being asked to repeat sentences, and this was immediately followed by a picture description task. It was found that descriptions tended to follow the primes in their structural features, so that a passive sentence would follow a passive prime, etc.

Levelt (1983,1989) analysed speakers' self-interrupting self-repairs, as in:

(2.18) From white I go straight to—er—right to blue.

Under normal circumstances such repairs—known as *reformulations*—have been found to conform to a *well-formedness-condition*: a repair is well-formed if a hypothetical continuation of the original is capable of standing in a well-formed disjunction along with the repair; for example:

(2.19) From white I go [straight to (green) or right to blue]

Reformulations often involve considerable retracing—additional evidence that surface forms are being referred to.

Complementary evidence for the short-term persistence of surface forms, and their use as a cohesive resource by language producers, comes from the study of anaphora and ellipsis. Sag and Hankamer (1984) studied the phenomenon of deep and surface anaphora. While the former category may be distinguished by the possibility of deictic interpretation, as in: "I wonder who she was", examples of the latter depend on surface-structural parallelism, and assume a surface antecedent for their completion, as in VP ellipsis:

(2.20) *C*: is flight 504 on time
A: it should be

Whereas deep anaphora may be explicated in terms of reference to entities in a discourse model, surface anaphors require an analysis in terms of surface structure. This distinction is supported by the observation that surface anaphora are characteristically limited in their textual scope.

Short term persistence may be more or less guaranteed, but there is evidence that longer term persistence may depend on the functional load placed on particular occurrences of surface forms. Bates *et al.* (1978) analysed conversations from television drama, and found better recall of expressions which explicitly introduced referents, compared with anaphoric or elliptical expressions. Johnson-Laird (1983) describes an experiment where subjects are required to recall descriptions of spatial configurations. Those descriptions which were indeterminate (capable of more than one interpretation) gave rise to better verbatim recall than determinate descriptions. Johnson-Laird accounts for this in terms of model building: in the case of

determinate descriptions, mental models are built or extended, so memory for gist is uppermost; in the case of indeterminate descriptions, subjects need to recall a propositional form—and the verbatim description is one way of doing this.

Evidence that speakers make use of repeated surface features in order to maintain conversational coherence comes from the analysis of Schenkein (1980), who applied the techniques of Conversation Analysis to transcripts of natural conversation. A major finding of this work was the number of conversations which exhibit repeating patterns of sequential organisation. The parallelism is reinforced by the re-use of both structural and thematic material:

(2.21)

2nd Voice: ... My eyes are like organ stops, mate...

.....

1st Voice: ... Cor, the noise downstairs, you've got to hear and witness it to realise how bad it is.

2nd Voice: You've got to experience exactly the same position as me mate, to understand how I feel. My eyes are so bad they are blurred and I've been using (binoculars) all night.

Not only do the speakers play off similar complaints against one another, but they employ parallel surface resources which reinforce this patterning. Nevertheless, despite the evidence for *repeated action sequences*—repeated sequences of argumentation extending over a number of turns (Schenkein gives an example of a repeated six-position sequence)—the most striking instances of structural parallelism, Schenkein concedes, occur within a single turn or a pair of adjacent utterances.

The use of surface descriptions as a resource in conversation has been studied experimentally by Clark and Wilkes-Gibbs (1986), whose subjects communicated about a set of geometrical figures, which required some inventiveness to describe. It was found that once a description was settled upon, subjects tended to stick to it. Garrod and Anderson (1987) frame the issue in terms of the resources available to communicating agents in achieving coordination of action; they may build common linguistic representations, and they may rely on shared internal models. In a series of experiments in which dialogue partners talked one another through a computer "maze game", Garrod and Anderson found that pairs of players evolved their own protocols for describing spatial configurations. These protocols were interesting not only for their surface regularity, but for regularity according to the type of model

they presupposed. Garrod and Anderson found the stability of these 'description schemes' to be achieved not so much by explicit negotiation, as by an iterative process which they call *input–output coordination*: a speaker formulates descriptions with respect to a model which he assumes to be held in common with the interlocutor; in interpreting the other's utterances, such assumptions may be incrementally revised until stability is reached. They call such a strategy "falsification definite", since it proceeds on the assumption that convergence exists, until evidence suggests otherwise. Similarly, Clark and Brennan (1991) describe how speakers attempt to establish common ground with a minimum of collaborative effort. This process of *grounding* takes place via protocols which may include the use of verbatim repeat in backchannel signals.

2.3.3 Summary

There is a considerable convergence between researchers from different fields about the existence of an internal *discourse model*, which a conversing agent accesses and updates during language processing, and which provides a resource for continued conversation. In psycholinguistics, it has been posited to account for retention and accessibility of information. It has also been proposed that various forms of linguistic marking conventionally indicate to the interlocutor the degree of accessibility of a discourse referent. AI models of language behaviour make use of databases (or *knowledge bases*) as a necessary component to keep track of an agent's knowledge. Partitioning techniques have been used to represent focus and topic, and may be used to account for certain cases of anaphoric reference. The notion of the discourse model is also consistent with model-theoretic approaches to language meaning, though the latter often fail to distinguish clearly between mental models and the world that is being modelled. The representations used vary from sets of propositions (logical and some psycholinguistic models), via models which make relations explicit in an analogical manner, to semantic networks, favoured in computational implementations. It may be that such representations are weakly equivalent, in that one can be used to model any other. In any case, the notion of a knowledge base used as a resource

by an agent underlies them all.

However there is linguistic evidence that recent surface forms are treated anaphorically in a different manner to longer-term discourse referents; this is backed up by psycholinguistic accounts, which demonstrate that an agent does keep a short-term record of linguistic structures. According to the Input/Output model of Garrod and Anderson, the use of both short term and longer term records by agents serves pragmatically to reinforce the assumption of convergence. As I discuss in Section 2.5.2, the prosodic accentual properties of utterances also provide evidence for discourse models. It is possible that the distinction between deep and surface anaphora may also apply here.

2.4 A model of the speaker

2.4.1 Cognitive models of language production

The 'speech error' model of language production (eg. Garrett 1980) possesses, apart from the message level containing pre-linguistic representations, two levels: the *functional* and the *positional*. The former consists of meaning-based representations of lexical items, assigned to functional syntactic roles; their function is to control the elaboration of syntactic structure. At the positional level, phonologically explicit representations of lexical items are assigned linear positions, their function being to control the elaboration of phonetic form. Levelt (1989) discusses the architectural characteristics that a plausible cognitive model of production should possess. The principle of information encapsulation states that components should be relatively autonomous—that is, require a minimum of interaction with other components in order to obtain information, and operate on a *characteristic input*. Processes are thus specialists with regard to their input and their mode of operation. Levelt's architecture is shown in Figure 2.3. The model is based on the evidence of errors committed by speakers, notably word and sound exchanges. Levelt divides utterance production into three stages: conceptualising, formulating and articulating. The Conceptualizer is concerned with elaborating a *preverbal message*; the Formu-

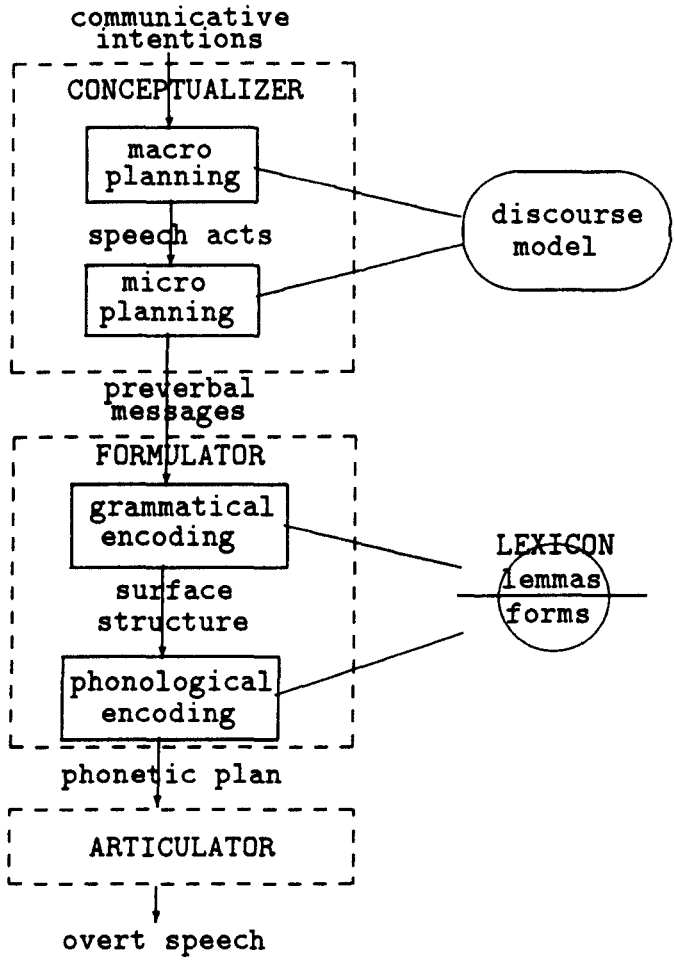


Figure 2.3: Levelt's 'blueprint for the speaker' (modified version)

lator takes this message as input and produces a phonetic plan which is used to drive articulatory processes. The specialist nature of the Conceptualizer and the Formulator is clear from the figure. The former makes use of data from the discourse model, while the latter uses the lexicon as its knowledge source.

The Conceptualizer may be split into two components: a macroplanner and a microplanner. The macroplanner elaborates communicative intentions into *speech act intentions*, where the latter are specified for content and syntactic mood. In the second stage, microplanning, issues such as thematic roles and accessibility marking are decided. The preverbal message produced

> "... is a semantic representation that is cast in the propositional language of thought but that, at the same time, meets conditions that make it expressible in natural language." (Levelt 1989:75)

The Formulator consists of a grammatical encoder, and a phonological encoder, the distinction between these being based on the same considerations as that between Garrett's functional and positional levels. Grammatical encoding takes as input the preverbal message, and produces a syntactic surface structure in which the terminals are abstract lexical entries, or *lemmas*, together with sufficient morphosyntactic featural information to specify any inflections and provide for detailed phonological encoding.

The architecture also provides for monitoring components (not shown in Figure 2.3). Levelt envisages a feedback loop whereby the phonetic plan is processed by the Speaker's comprehension system; a monitor within the conceptualiser then assesses the appropriateness of the output, and takes remedial action where necessary, giving cause to self-interruption and self-repairs. Such feedback is possible because the system is incremental. Autonomous processes work concurrently, and operate in an 'eager' fashion, as soon as some minimum amount of their characteristic input becomes available. These processing characteristics, it is generally accepted, must be required for human utterance production to operate efficiently from left-to-right, with the speaking of an utterance beginning before its planning is complete.

The issue of how much information of a contextual nature goes into an utterance however is not clear. Consider the case of phrasal utterances such as noun-phrase

answers, which are incomplete without reference to the question. Are these simply conceptualised as fillers for gaps, and passed to the formulator without any accompanying contextual evidence? It would appear not, if evidence from languages such as German, where an answer may exhibit case agreement with the environment of the question, is to be taken into account.

2.4.2 Natural language generation: computational models

The work of natural language generation programs is often divided between a *conceptual planning*, and a *linguistic realisation* stage. While these correspond reasonably well with the conceptualisation and formulation stages of the levels architecture, they are principally motivated by a need to separate out conceptual and linguistic knowledge bases, and the processors special to each.

2.4.2.1 Planning utterances

Conceptual planning—sometimes referred to as *strategic generation*—was investigated by McKeown (1985), whose generation system TEXT is able to describe information and meta-information from a military database. At the top level, paragraphs are planned on the basis of *rhetorical schemas* such as "definition", "compare-and-contrast". These schemas are then instantiated by information from the database, selected according to principles of topic coherence, and continuity of local focus.

A different approach was taken by Appelt (1985), whose utterance-planning component is based on Cohen's account of Speech Acts as plan operators (cf. Section 2.2.1). Given a high-level intention, such as to get an apprentice to do something with a screwdriver, the planner is capable of reasoning about knowledge and intentions of discourse agents, in order to derive the most appropriate utterance. The possibility of plan optimisation means that Speech Act operators may be combined into a single utterance. For example an agent may combine the goal of getting the interlocutor to pick up an object with that of informing the interlocutor where the object is, to produce:

(2.22) Get the screwdriver from the floor behind the table

Appelt incorporates the conceptualisation of referring expressions into plans, using the notion of a *referring action*, whose perlocutionary effect is achieved if the hearer becomes aware of the relationship between the intended referent and the noun phrase description used. Because noun phrases normally do not appear in isolation, referring acts will normally be combined with other acts—as in (2.22), where the act referring to the position of the screwdriver is combined with the instruction to get it.

A great deal of the work on planning for natural language generation has been concerned with planning the presentation of information, as in the case of Appelt's and McKeown's work. Hovy (1990) describes a program which, he claims, is sensitive to the interpersonal relationship, actual or desired, between speaker and interlocutor. The program is capable of describing information from the same knowledge base according to different sets of criteria, such as the speaker's intention to influence the hearer, or the hearer's perceived emotional state. On the basis of a given set of criteria, *rhetorical goals* defined in terms of desired stylistic properties of the text are set up. Thus the rhetorical goal *formality* is given the value *colloquial* if the depth of acquaintance is that between friends, but it then receives a more formal value if in addition the effect of distance is required. Rhetorical goals act as global parameter settings which guide choices such as depth of expansion during planning or selection of appropriate vocabulary. Conflict between rhetorical goals is handled not explicitly, but by averaging their respective performances in terms of the number of choices they have influenced. Another aspect of planning, that of referential expressions in discourse, has received attention in Dale (1990). In Dale's generator, descriptions vary in detail according to the amount of information needed to effect discrimination of referents; decisions to use pronouns or zero-anaphora are based on recency considerations.

2.4.2.2 Linguistic realisation

This stage, sometimes known as the *tactical* stage in language generation, takes place once a suitably tailored conceptual representation has been produced. Constraints defined in a grammar and/or lexicon are then applied to produce natural language

output. It is possible to identify two major strands of work which have been active in recent years. Computational linguists concerned with elegance and economy of representation have built models which make use of constraints defined in a *bidirectional* grammar/lexicon—ie, one in which knowledge is declarative and thus indifferent to how it is to be applied, whether in parsing or generation. Both generation and parsing with such linguistic representations may be viewed as the process of deriving an *analysis tree*, which is equivalent to a proof of the well-formedness of the input, be it conceptual or textual (Shieber 1988). The nodes the analysis tree are typically labelled with feature structures that combine lexical, syntactic and semantic properties. Constraints are combined by unification. There are of course a number of search regimes that will lead to the same result, some more efficient or complete than others. Of particular interest is the *head-driven bottom-up algorithm* (Shieber *et al.* 1990, van Noord 1990). In this approach, search is lexically based: lexical items are selected according to their ability to match with the incoming semantic structure; if a lexical item carries with it constraints on its linguistic environment, these are used to provoke further search.

The other continuing major strand of research concerns attempts to provide computational models of theories of incremental production. Kempen and Hoenkamp (1987) describe a model of production where the generation of syntactic forms is supervised by concurrent grammatical 'specialists', corresponding to the major constituents of surface structure.

POPEL-HOW (Reithinger 1989, Finkler and Neumann 1989) represents a coming together of several strands of generation research. Firstly, it is an incremental generator. Output is available at the phonetic (word string) level before the conceptualisation component has finished. Reithinger *et al.* demonstrate that this results in word order variations reflecting different orderings of conceptual input; given a destination (Detroit), a travelling action, and a traveller (Peter), in that order, the formulator produces (2.23); If the vehicle (airplane) is then given, the continuation (2.24) is produced. On the other hand, specifying the vehicle, the person, the action and the destination in that order, produces (2.25):

(2.23) Nach Detroit fährt Peter
(*To Detroit travels Peter*)

(2.24) mit dem Flugzeug
(*by plane*)

(2.25) Peter fliegt nach Detroit
(*Peter flies to Detroit*)

It is thus possible to account for word-order variations in terms of the order in which the message is put together, which in turn may reflect topicality.

Secondly, feedback channels between levels are permitted. If the Formulator cannot continue because the linguistic structure it is elaborating requires more conceptual structure, this can be sought at the conceptual level. Finally, POPEL-HOW employs a declarative, lexicon-based unification grammar, with the virtues of modularity and maintainability that this entails. Incrementality with feedback is achieved by the procedural means of communicating concurrent objects; however the linguistic constraints which these objects are forced to obey are compiled from the grammar.

2.4.3 Summary

There is general agreement between cognitive modelling accounts of language production and computational implementations, concerning the separation between the conceptualising and formulating stages. Many of these are based on the notion of *hierarchical planning*, whereby structures conceived at a higher level are refined and elaborated at the lower levels, with the aid of context-independent knowledge bases. Hovy (1990) however raises the issue of *restrictive planning*, which he claims is needed to account for the phenomena of his model, such as the balance of possibly competing stylistic goals. For such a model to work, it is necessary that local execution choices at a number of levels are monitored with respect to their effects in satisfying persistent rhetorical goals, and their outcome is adjusted so as to contribute towards global satisfaction of these goals. Such a processing model may be particularly suited to describing how prosodic choices are made during production.

2.5 The relevance of prosody

The study of intonation, as Hirst (1987) points out, cuts across traditional subject boundaries, "from phonetics to pragmatics and far beyond". In this review, I shall be concerned with two aspects of prosody: accent and melody, and with what pragmatic factors affect them.

2.5.1 Prosodic focus and accent

The phenomenon of sentence accent is one that has long absorbed analysts. Unlike lexical stress, which is more or less stable, in a language like English more than one accentual pattern may apply to a single sentence:

(2.26) a. The average American expécts too much of people.
b. The average American expects tóo much of people.
c. The average American expects too múch of people.
d. The average American expects too much óf people.
e. The average American expects too much of péople.

(Bolinger 1989: 363)

A number of explanations of sentence accent have been attempted, from relatively syntactic to relatively semantic/discoursal. They have in common an appeal to organisation at a level beyond that of the word.

The notion of Figure and Background may have found its way into linguistics from Gestalt Psychology via the Prague School (Danes 1960), where it was considered to be a major factor underlying the organisation of material within the sentence. This was particularly appropriate to a relatively free-word-order language like Czech, and in the guise of the theme-rheme opposition it has been applied with some limited success to English. The terms *focus* and *presupposition* were applied by Chomsky (1971) as part of an attempt to incorporate phenomena of accent placement into the then prevailing methodologies of Generative Phonology and Transformational Syntax. Chomsky, and following him Jackendoff (1972), were able to provide a meaning representation for focus and presupposition, in terms of a variable extracted from logical form. Chomsky proposed that a single accent could give rise to a number of interpretations, in terms of the domains it brought into focus. For example, the

sentence (2.27) with accent on the final word may be construed as an answer to any one of (2.28–2.30).

(2.27) Papa has given Tommy a GUN
(2.28) What's happened?
(2.29) What has Papa done?
(2.30) What's Papa given to Tommy? (Gussenhoven 1987: 12)

Similarly, Wilson and Sperber (1979) point out that final accent on a sentence such as *You've eaten all my apples* is ambiguous as to which of presupposition is intended, for example among: *you've eaten all of something, you've eaten something, you've done something, something's happened.* Sperber and Wilson refer to these as a sequence of *ordered entailments*, because earlier items logically entail later ones. They suggest that the ambiguity as to which entailment is intended can be utilised by listeners, who make successively more detailed hypotheses, during left-to-right processing, corresponding (in reverse order) to the set of ordered entailments.

Focus thus has been seen by many analysts as a binary distinction, albeit one whose scope with respect to an observable pattern of sentence accents is ill-defined. Many prosodic phonologists have taken the basic notion for granted, concentrating instead on the factors which govern accent placement, given that focus is somehow independently assigned. The original Generative Phonology position (Chomsky and Halle 1968) was that according to the Nuclear Stress Rule the major accent was placed on the final accentable syllable of the sentence. But there were exceptions. Schmerling (1976) examined a body of data and found that in many cases 'news sentences' which seemed not to depend on given material, did not obey this pattern.

(2.31) Johnson died

These examples Schmerling accounted for with the principle that semantic predicates (and hence verbs) were semantically subordinate to their arguments (typically noun phrases). The SVO configuration typical of English, Schmerling claimed, had tended to obscure this fact, which became apparent when intransitive and phrasal verb

examples, or examples from languages such as German with different basic word order patterns, were considered.

Selkirk (1984) and Gussenhoven (1983) both endorsed Schmerling's approach. Gussenhoven, starting with the assumption that the binary-valued feature *focus* "exists as a formal category available in speakers' grammars" (Gussenhoven et al. 1987: 4), is concerned with predicting where accents will fall, given a distribution of the feature over a sentence. Making use of the notion of *focus domain*, defined as

> one or more constituents whose [+ *focus*] status can be signalled by a single accent (*ibid*: 15)

Gussenhoven proposes a system of rules which effectively merge and redistribute focus into domains from which accent placement is easily accounted for. An example of these *Sentence Accent Assignment Rules* (SAAR) is given in (2.32).

(2.32) Domain assignment: $\underline{P}(X)\underline{A} \rightarrow [P(X)A]$ (Gussenhoven 1987: 16)

The rule states that a focussed predicate, followed by some non-specific unfocussed constituent, followed by an argument, may be combined into a single focus domain. Schmerling's rule, that arguments rather than predicates were accented, then applies. Gussenhoven's rules are fairly elaborate, designed to minimise the number of exceptions. They have been applied to a sizable body of data for both English and Dutch. Experimental studies (Gussenhoven 1984b) have added support. For example, when the condition of (2.32) is not met, for example when X is focussed, Gussenhoven found that a significant proportion of listeners were able to detect more than one accent; this corresponds (according to Gussenhoven) to the case in which 'focus domain merging' is blocked.

Against all the work on focus domains must be set the position of Bolinger, who has consistently opposed attempts to give structural accounts of accent placement (Bolinger 1972, Bolinger 1985, Bolinger 1986, Bolinger 1989). According to Bolinger,

> Accents are prima facie iconic, responding to the speaker's sensation of the INTEREST in what he is saying ... At a first remove from interest we have IMPORTANCE—what is most important is what is apt to be most interesting; and at a second remove we have INFORMATION—what is

most informative is apt to be what is most important. Information structure is served by accentual structure, but mediated through sensation. (Gussenhoven et al. 1987: 95)

Bolinger points out, for example, that the often-observed de-focussing of verbs in news sentences can be accounted for in terms of sacrifice of accent to a nearby focal meaning, and because the average noun is more interesting than the average verb. When a verb is interesting, it is quite normal to accent it:

(2.33) What happened?— Tommy **saw** a bear!
Tommy spótted a bear!
Tommy **caught** a bear!
Tommy cáptured a bear! (Bolinger 1986:125)

The verbs *see* and *catch* are relatively commonplace and uninteresting; *spot* and *capture* are less frequent, and semantically more specific; they are more likely to be accented. By the same token, nouns which are relatively semantically empty, such as *people* may lose their accent to the verb—compare (a) and (b) below:

(2.34) a. Sóldiers were standing guard
b. People were standing guárd (Bolinger 1989: 235)

As a complement to the principle of interest in determining accent placement, Bolinger proposes the additional principle of *power*. He suggests that *ceteris paribus* a pattern in which "a sort of annuciatory or attention-getting accent comes towards the beginning and a 'punch' accent comes towards the end"—ie, an early onset and a late nucleus, to use Crystal's terminology—is inherently more powerful in its effect than one in which the the nucleus is early; compare:

(2.35) a. We'll sée about that.
b. We'll see about thát. (Bolinger 1986:82)

In fact the shifting of accents as far to the right as possible (causing at times distortions in the default position of lexical stress) allows for the possibility of more prenuclear accents, and thus a greater effect of power:

(2.36) a. I don't WATCH television!
b. I dón't watch TELevision!
c. I dón't watch teleVISion!

d. I DON'T WATCH TELeVIsion!

(Bolinger 1989: 377)

The sequence in (2.36) corresponds to increasing emotive pressure on the speaker's behalf, occurring perhaps because of misunderstanding or communication failure.

2.5.2 Prosodic marking of accessibility

We have seen (Section 2.3.1) that a number of linguistic devices such as anaphora can be used to indicate accessibility in the assumed shared discourse model. The notion of 'Given' and 'New' was initially used by Halliday to account for patterns of prosodic accent. Subsequent investigations of the use of accent, and its correlation with accessibility, have added to our understanding of the Given-New distinction. Brown (1983) applied the classification of Prince (1981) to the analysis of the results of an experiment in which speakers described figures containing simple shapes. She found, as would be expected, that evoked items tended to be deaccented, and new items to receive accent. Interestingly, material not directly mentioned, but inferrable from the preceding discourse, also tended to be accented, though slightly less frequently.

Terken (1985) carried out a series of experiments in which speakers described the relative positions of letters on a simple VDU display, using sentences such as:

(2.37) The P is below the K

Because the sequence of configurations was so arranged that there was continuity over time, he was able to analyse both cumulative mention, and probability of mention of an item, as well as the 'pragmatic status' of an item (whether it was currently considered to be movable or stationary). His results showed that items tended to be treated as given—ie, to be unaccented—relative to previously mentioned items, when these were either: (i) coreferential, (ii) had the same pragmatic function (in the sense that it belonged to the same class, of either movable, or stationary items; (iii) had the same sentential position. Nevertheless, even strongly predictable items lost their propensity to being defocussed if in the subject or predicate position where that position in the previous utterance was filled by another referent. In a further experiment, in which speakers gave oral instructions concerning a visual task, Terken

found that the boundaries between instructions inhibited carry-over of what was accessible, apart from instruction topics and topical items. This result is strongly reminiscent of those of Sanford and Garrod (1981) on scenario-dependent entities (cf. discussion on page 31).

The accentual pattern of an utterance may act as a positive aid to the listener. Non-accented syllables tend to be phonetically reduced; for accented syllables the converse holds. Cutler and Fodor (1979) suggest that the comprehension of accented words is faster; this allows listeners to give priority to the processing of new information in an utterance. Fowler and Housum (1987) presented listeners with materials obtained by using various spliced combinations involving first and subsequent mentions of words, from spontaneously spoken monologues. They found that the subsequent mentions were perceptually degraded, but that listeners were able to compensate because they had heard the words before, and because by the time of the second mention there was usually added contextual support. They also found that listeners use their knowledge that a word has been repeated to facilitate recall of its prior context. Terken and Nooteboom (1988) present results with similar implications. Subjects followed successive configurations of letters on a visual display, similar to that of (Terken 1985), and were given the task of verifying recorded utterances. Reaction times were measured, with the clear result that not only was new information verified faster when the information-carrying words were accented than when they were not, but given information was verified faster for unaccented words, than for accented. These results suggest, as Terken and Nooteboom hypothesize, that both accentuation and de-accentuation are exploited by listeners: accentuation directing the attention of the listener to acoustic/phonetic decoding (ie, bottom-up processing); de-accentuation leading him to search among the limited set of discourse entity candidates that are currently active, thus allowing top-down processing.

Further evidence for the usefulness of prosodic focus to listeners comes from an experimental study by Blutner and Sommer (1988). When subjects were presented with contextualised ambiguous words, initially both readings were present, if the words were part of the semantic focus of the sentences. For words not belonging to the semantic focus (ie, material which could be considered 'given' or presupposed),

there was not found to be such a stage of activation preceding disambiguation.

These findings complement Ariel's (1989) and Givón's (1992) conjectures about the use of explicit high-accessibility markers such as pronouns, in indicating to the listener how to retrieve a previous referent. In the case of unaccented forms, these were not attenuated at a lexical or semantic level; however, as Fowler and Housum verified, they tended to be prosodically attenuated. Givenness marked by prosodic attenuation appears therefore to facilitate search for listeners.

2.5.3 Intonation and contrast

The idea that contrast can be marked by accent must be uppermost among folk-notions of prosodic function among English speakers. Contrastive accent has even found its way into written texts, via the typographical device of italics. Nevertheless a precise linguistic characterisation of contrast has proved difficult. Objecting to the classification of some pitch contours as contrastive or emphatic, Bolinger (1961) retorts:

> As far as we can tell from the behaviour of pitch, nothing is uniquely contrastive.

As support for this view Bolinger gives the examples:

(2.38) [↑He didn't buy a **LH** Ford] he bought a ↑**HL** Plymouth.
(2.39) [↑Just leave him **LH** alone] and he won't ↑**HL** bother you.

Whereas some kind of contrast is undoubtedly being made in (2.38), the same contour is used in (2.39) without any such intended effect. So far as *semantic* contrast is concerned, Bolinger (1961) points out that at one extreme, any item that receives pitch prominence may said to be in contrast, even if the class of potentially contrasting items is hopelessly broad—as is the case in *Let's have a picnic*, where "picnic" may be taken to be in opposition with any of the innumerable other things we might do.

Chafe (1976) on the other hand, asserts that "contrastive sentences are qualitatively different from those that simply supply new information from an unlimited

set of possibilities". What distinguishes the truly contrastive case, he claims, is an awareness on the speaker's part that some item is being selected from a limited set of candidates, and it is this that is the correct one. Chafe further suggests that some contours may be intonationally distinguished as contrastive:

(2.40) a. They elected **H** Alice **HL** president
b. They elected **HLH** Alice **HL** president

In (b) for example, the speaker is consciously contraposing Alice with other names on the committee; this is not the case in (a).

For Ladd (1980) what appears phonetically contrastive may be accounted for in terms of *narrow focus*. De-focussing of 'given' material leads to its being *deaccented*, with the result that unlikely elements may appear to achieve contrastive prominence. Thus in (2.41:B) the nucleus is on *read*, because *books* is deaccented:

(2.41) A: Has John read Slaughterhouse-five?
B: No, John doesn't READ books

An interpretation where the second *read* is contrastive is clearly not tenable.

Ladd also drew attention to a class of cases where a fall-rise nucleus is used to indicate *focus within a given set*:

(2.42) A: Did you feed the animals
B: I fed the **HLH** CAT

Ward and Hirschberg (1985) follow up this analysis, pointing out that in many cases a set-theoretic explanation is too limited. They propose instead that the fall-rise in these cases introduces the pragmatic implicature of *uncertainty*; the speaker being uncertain whether his/her offering will meet the expectations of the interlocutor.

Couper-Kuhlen (1984) proposes two basic semantic types of contrast. Like Ladd, she does this in truth-conditional terms. In semantic type 1, associated with a ↑**HL** contour,

> "the speaker *asserts* that a proposition (or an item in a proposition) is true and simultaneously *asserts* that a contrasting proposition (or a contrastive item) is false.

Thus *We're going to* ↑**HL** *PORTland* denies the possibility of some other activity or destination. Couper-Kuhlen's semantic type 2 corresponds essentially to the use of

the fall-rise studied by Ladd and Ward and Hirschberg. Here the speaker is being concessive rather than assertive. Cutler and Isard (1980) likewise maintain that contrastiveness is signalled intonationally. On the basis of the example:

(2.43) a) London's the capital of Scotland isn't it?
b) No **HL** Edinburgh's the capital of **HLH** Scotland]
HLH London's the capital of **HL** England

Cutler and Isard claim that intonation can distinguish items which are contrasted from those focussed items which are not.

What none of the above analysts have paid particular attention to is the frequent co-occurrence of contrastive prosody with situations of interactive repair. Levelt and Cutler (1983) analysed repair exchanges from a corpus in which speakers described visual patterns. They found that in the case of *marked* repairs—ones where the prosody differed from that of the original—the difference between the reparandum and the repair was predominantly semantic. Moreover, there was a significant correlation between the markedness and the size of the semantic field from which contrastive items were taken. Thus repairs involving direction (only four directions possible) were more likely to be marked than those involving colour (twelve possibilities).

The evidence points against the existence of intonational forms employed uniquely to express contrast. Nevertheless, with more sensitive definitions of semantic contrast, it may be possible to formulate general rules for contrastive intonation. It may be noted that many examples illustrating contrastive intonation refer to disparate states of affairs. In Section 4.3.4 I examine the contrastive utterances that get produced when such conditions hold, and the intonational forms that may be associated with them.

2.5.4 Intonational contour and meaning

Ladd (1980) likens the task of attempting to assign meaning to intonational units to that of the extra-terrestial linguist confronted with the phrases *sitting in a chair* and *sitting on a chair.* To arrive at the (correct) insight that the two expressions have different semantic implications requires first the knowledge of phonetic detail

and phonological structure which will enable them to be told apart. In the case of intonation, although broad agreement exists with regard to the phenomenon of pitch accent (see Introduction: 1.4.2), there is considerably less agreement about structural phenomena than that existing for segmental phonology. One relatively uncontroversial notion is that the phenomenon of contour as melody, and the phenomenon of accent are to some degree separable, at least where function is concerned. We have seen in the previous sections that an interpretation of accent as applying to information status is perfectly maintainable. In fact the results of Cutler and Swinney (1987) demonstrating that young children are unable to make use of accentual cues to focus, when set against their known competence in using melody pragmatically, suggests that these functions are also distinct developmentally. This section is devoted to the melodic aspect of intonation, and its possible pragmatic interpretations.

Within linguistic approaches to intonational meaning, there has been a strong tendency to take as units of functional analysis those defined on independent formal grounds as phonological units. In the 'British tradition' (eg. O'Connor and Arnold 1961) the intonation phrase, or tone group, is organised as follows:

PREHEAD HEAD NUCLEUS TAIL

where the nucleus is the location of the final—and usually perceptually most prominent—pitch movement; the head is that portion from the first pitch accent to the nucleus, and the prehead and tail consist of unaccented syllables, though in the case of the tail these follow the contour established by the choice of nuclear tone. This terminology has undeniable descriptive advantages. Analyses based on it however tend to equate a homogeneous set of form options with an equally homogeneous set of functions. Thus Gussenhoven (1983) considers the options of fall, fallrise, rise to be equivalent to various manipulations of the 'background' with respect to the 'variable' (where these terms correspond roughly to the Chomskyan 'presupposition' and 'focus'):

Adding the variable to the background	*fall*
selecting the variable from the background	*fallrise*
testing the validity of the variable with respect to the background	*rise*

Gussenhoven claims that the use of the nuclear tone paradigm can be accounted for in terms of these manipulations.

In those analyses which isolate the nucleus as the most meaningful unit, there is a tendency to apply circular reasoning in determining what choices of nuclear unit there should be. Thus Ladd (1980) distinguishes between the high and low rise—making no such distinction for falls—largely on functional grounds, and Brazil *et al.* (1980) are able to collapse all nuclear options into two, the 'proclaiming tone' (typified by a falling nucleus) and the 'referring tone' (typified by the fall-rise). In this class of models, there is also a tendency to consider the prenuclear part as subsidiary. Thus Levelt (1989):

> "The intonational meaning of the phrase is essentially carried by the nuclear tone. The prenuclear tune can modify that meaning—can soften it or sharpen it—but cannot essentially change it.

In contrast, the autosegmental approach of Pierrehumbert (1980) lends itself to a treatment in which the tonal units, and even their components, are assigned meaning. Pierrehumbert's phonology has been reviewed in Section 1.4.2. In Pierrehumbert and Hirschberg's (1990) account of intonational meaning, the phonology is mildly hierarchical, consisting of *pitch accents*, *intermediate phrases*, and *intonational phrases*. Phrase accents and boundary tones are taken to mark the end boundaries of the latter two categories. The model of intonational meaning then proposes that pitch accents convey information about individual discourse objects. Thus **H*** indicates that the salient object should be treated as 'new', whereas **L*** is taken to mean that the salient object should be excluded from the predication of the utterance, so that in:

(2.44) **L*H$^-$H$^\%$** I should apologize

the speaker is declining to commit himself to apologize, though it might be inferred that the addressee believes it should be the case. Pierrehumbert and Hirschberg's account extends to the meaning of phrase accents and boundary tones, which are seen as relational markers to the surrounding discourse. They claim that their fine-grained approach allows more generalisations than a nuclear-tone based one such as

Gussenhoven's. The Pierrehumbert-Hirschberg theory of intonational meaning has been further endorsed by Hobbs (1990).

Opposed to such *compositional* approaches to intonational meaning have been attempts to analyse intonation in terms of holistic units with their own specific functions. Liberman and Sag (1974) claim a correlation between such 'holistic contours', and illocutionary forces, and posit the existence of an *intonational lexicon* where such correspondences are defined. Sag and Liberman (1975) make the narrower claim that certain contours are capable of freezing the illocutionary interpretation of an utterance, so that only the direct speech act reading became available. Thus in:

(2.45) $_1$Why don't you move to Cali $_2$fornia

a $_0[^{\downarrow}$ $_2$**HL** reading was capable of being interpreted as a suggestion, whereas the 'tilde contour': $_1$**HL** $_2$**LH**, they claim, can only have the direct reading of a question. Cutler (1977) however points out that the contours proposed by Liberman and Sag were far from being restricted to the illocutionary meanings intended for them. Liberman and Sag's 'contradiction contour', for example, could be used in a situation where no contradiction, only disapproval, was present:

(2.46) **HL** Go and see what the fellow **LH** wants

Cutler's conclusion, that intonational meaning cannot be decontextualised in this way, is in line with Bolinger's scepticism (cf. page 52) towards claims for contrastive contours.

Faced with the analytic difficulties exemplified above, a functional approach could be considered more appropriate to this study, especially if it is the case that within a limited domain, those functions may be enumerated. In the remainder of this section therefore I review a number of those functions which are relevant to this work, and which have been thought to be influenced by intonation. The case of the use of contour in contrastive intonation has been reviewed in Section 2.5.3.

Prosodic signalling of turn-taking One aspect of turn-taking that continues to puzzle researchers is how synchronisation of turn-change happens so efficiently.

Sacks *et al.*(1974) propose that linguistically-definable *transition relevance places* exist in a speaker's utterance, and that listeners planning to take the floor project the position of these so as to know when to intervene. One obvious cue to such potential turn boundaries is sentential completeness; however a speaker may choose to draw out a sentence, adding postmodifiers after minimal completeness has been reached. Prosodic cues seem a likely possibility, especially taking into account the results of experiments such as that of Cutler (1976) which suggest that listeners monitor the progression of the intonation contour and are able to predict prosodic events later in the utterance. They are also attractive to phonological accounts of intonation which isolate the final boundary tone as a unit of analysis, if this can be shown to function as such a cue.

Duncan (1974) first advanced the view that intonation acted as one among a number of turn-giving cues.[2] According to Duncan's analysis of videoed conversations, most final pitch movements acted thus as 'turn-taking signals'. Beattie, Cutler and Pearson (1982) investigated television interviews with the British politician Mrs Thatcher. They found that interruptions by the interviewer could be consistently attributed to lack of consistency in turn-giving signals. Cutler and Pearson (1986) got subjects to read dialogues where the same target sentence was used either turn-medially or turn-finally. They failed to find a significant correlation between position in the turn, and the intonation used. Likewise, listeners did not prove to be competent judges of whether the sentences presented in isolation were turn-medial or turn-final. However there was some consistency between the actual contours used by speakers, and listener's judgements: upstepped contours tended to be judged turn-medial, while downstepped contours were pronounced turn-final.

Somewhat less conclusive are the results of Schaffer (1983). Listeners were asked to judge whether excerpts from recorded dialogues were turn-initial, final or medial. Where lexical/syntactic cues were present, listeners appeared to pay more attention to these than to prosodic cues. Despite listeners' lack of consistency and correctness in interpreting these, Schaffer found that rises were more successful cues to turn-

[2]Other cues included for example when the speaker turned his gaze to the interlocutor.

finality than falls. Schaffer speculates that local cues such as contours may be of less importance to speakers and listeners than their common grasp of the conversational organisation.

There is some evidence that heightened register and amplitude are used in cases of successful (as opposed to unsuccessful) turn-grabbing (see French and Local 1986).

Intonation and questions Correctly interpreting questions is clearly important for conversational participants, especially in information dialogues. It is well-known that syntactic markers such as sentential mood or the presence of tags are not in themselves sufficient to indicate whether or not an utterance is a question (eg. Quirk et al. 1985: 803–853). Can intonation help in this respect, especially when syntactic evidence is not present? Minimal pairs might suggest that this is the case:

(2.47) Yes-no question vs. exclamation

a. isn't he **LH** sure of himself
b. isn't he **HL** sure of himself (Couper-Kuhlen 1986: 148)

(2.48) Statement vs. question

a. **HL** John has
b. **LH** John has (Halliday 1967: 41)

But linguists have on the whole been wary about the existence of such a phenomenon as question intonation. Bolinger (1989) considers in turn a number of question types, asking what the correlation might be between those types and intonation contours. He concludes:

> One can calculate probabilities, but there are no defining connections between intonation and question type. (Bolinger 1989: 143)

He nevertheless is able to isolate cases where certain contours are typical or atypical of a particular question type. Thus a 'B + B' contour (in our terminology, ↑**HH**) predisposes an utterance to be a yes-no question; the contour '...(B +) B + A' (ie, [**HH**]^1 **HL**) is most typical of an alternative question; while *reprise questions*

(those that involve a more or less verbatim repetition of what has been said) 'strongly favour profile B'.

Brown, Currie and Kenworthy (1980) report work in which listeners were presented with question and answer utterances extracted from recorded natural conversations, either isolated or with context. The material was selected so that textual clues were not present. For the isolated utterances they found a good consensus among judges that a terminal high-rise indicated a question. Otherwise, listeners tended to agree but wrongly, judging the majority of questions to be not-questions. For contextualised utterances, the number of correct judgements was higher, as would be expected. Brown *et al* also examined the recorded corpus for correlations between question-type and contour. They found that polar questions associated with fall-to-low (**HL** $^{\downarrow}$]) were overwhelmingly *conducive*—ie, indicating to the addressee that a certain answer was expected. However the relevance of these findings to the work of this thesis may be limited by the fact that the data used is confined to the Edinburgh dialect of English.

Geluykens (1987) presented listeners with synthesized utterances, where the nuclear contours were varied according to the patterns laid down in Halliday (1970). The task was to grade the utterances, which carried no syntactic clues, on a scale from "definitely a question" to "definitely a statement". There was no reliable correlation found between question judgements and contours, though Halliday's tone 2 (high rise or fall-rise) did better than the others. The factor that did have a significant effect on listener's judgements was a pragmatic one. The grammatical subjects of Geluyken's sentences were varied, taking as values the pronouns *I, you* and *he*. An average of 53% of the 'you' utterances were taken to be questions, as compared with 12% and 19% respectively for the 'I' and 'he' ones. Thus *you feel ill* was more likely to be judged a question than *I feel ill*, irrespective of intonation. Geluykens suggests a Searlian interpretation (cf. Section 2.2.1): the 'you' utterances are more likely to fulfill the felicity condition, namely that the addressee can be expected to know the answer.

The affective use of intonation Bolinger (1986) points out that intonation in conversation typically forms part of a 'gestural complex' of signals which may include eyebrow-raising or the shrugging of shoulders. These gestural displays, it has been argued, have their origin in the emotional state of the speaker. But, Bolinger concedes, they are nevertheless susceptible to ritualization:

> "A ritual is arbitrary to the extent that the performer does not sincerely 'feel' the message he conveys. And yet the ritual is still close enough to the erstwhile reality it enacts so that it cannot be understood without reference to that reality. That appears to be the stage at which the supposedly arbitrary uses of intonation and its gestural counterparts have arrived. (Bolinger 1986:198)

Similarly, other accounts emphasise the universal symbolism of pitch. Ohala (1983) relates high pitches to displays of nonagressiveness or defensiveness. Conversely, low pitches tend to be associated with dominance and power. Bolinger (1986: 219) suggests that the use of falsetto register may have its origin in the display of submission; a falsetto terminal can be "unaggressively appealing".

Couper-Kuhlen (1986:185–7) makes the distinction between *emotion* and *attitude*. The former term, which covers descriptions such as *amused, angry, anxious, bored, frightened*, has to do with physiological arousal; whereas the latter is more descriptive of the speaker's overt behaviour: *affectionate, arrogant, coquettish, critical, deferential.* Bolinger's notion of 'accent of power' may be related to this attitudinal use of language. (cf. discussion on page 49).

Scherer *et al.* (1984) set out to investigate the varying assumptions of what they call the *configurational* and the *covariance* models. According to the former, linguistically-oriented hypothesis, both verbal and nonverbal features are category-valued; the choice of configuration of these produces different affective meanings. According to the covariance model, there is a directly observable covariance between the strength of speaker arousal and various acoustic parameters, such as pitch and loudness. Listeners were presented with recordings of utterances which were known to produce judgements of speaker affect. When the verbal component of the utterances was filtered out, it was found that affective judgements remained, thus confirming in part the covariance model. In a related experiment (Ladd *et al.*

1985) it was found that attributions of arousal appeared to vary continuously with the pitch range at which utterances were synthesized. However the 1984 results also provided some evidence for the configurational model, in that the interaction of contour type and verbal content was found to have an affect on judgements of affect.

2.5.5 Summary

Information focus and prosodic accent are related, albeit indirectly. Several researches have attempted to mediate this relation, using for example the notion of semantic predicate-argument structure, and broad/narrow focus. According to Bolinger however, sentence accents are primarily iconic, enabling a speaker to highlight those portions of an utterance which he deems interesting. Accenting is also exploited by listeners, as complementary psycholinguistic studies have shown, enabling them to direct processing resources to where they are most needed.

There has been a certain amount of confusion regarding contrastive intonation. There is for example little clear evidence that a single class of 'contrastive accents' exists; instead, examples from the literature have referred to a number of distinct phenomena. What these seem to share is a certain markedness, when compared for example with neutral accents.

A number of functions have been proposed the melodic component of intonation; these relate principally to the interpersonal aspect of communication. While the evidence concerning the use of intonation to signal question status and turn-taking is mixed, other studies point to the affective, attitudinal origin of prosodic contours.

2.6 Conclusions

This review began with an examination of theories of dialogue, both computational and non-computational. A number of theorists have chosen to emphasize the static effects of dialogue, as opposed to its dynamic behaviour. Speech Act based accounts, for example, are largely concerned with accounting for the effect of an utterance on the beliefs and goals of speaker and addressee. Computational accounts, especially

those pioneered by Cohen, Allen and Power, have contributed to a better definition and understanding of Speech Act theory. Other researchers (Sinclair and Coulthard, Houghton, Moeschler) who have appreciated the need for a structural account with wider scope than that of the isolated speech events have focussed primarily on the well-formedness properties of that structure. By contrast, Conversation Analysts have provided detailed studies of how speakers faced with online communication problems resolve these, or project their solutions further into the dialogue. However, their 'theoretical asceticism' (Levinson 1983), although healthy as a motivating force in empirical work, has on the whole prevented absorption of their ideas into cognitive and computational accounts.

A number of conclusions may be drawn from research which has concentrated on the internal representations which speakers and listeners may make use of. Linguistic and psycholinguistic studies both point to there being an essential difference between those *surface* representations built up during parsing and those residing in discourse memory. The latter are more permanent than the former; however the evidence is that within a limited time-frame language processing is capable of making use of recently generated surface representations. Many studies of discourse representation have been concerned with the resolution of anaphoric reference (eg. Sanford and Garrod 1981). Out of this concern have arisen computational models in which discourse memory is partitioned according to some criterion of focus or accessibility. Other studies have focussed on how mutual acceptability of discourse information is negotiated and how speakers use linguistic resources to explicitly indicate accessibility A great portion of this work however has focussed on the agent as understander rather than producer of language.

A 'levels-of-processing' model of language production with at least the stages: conceptualisation, formulation and phonological encoding, has long been accepted in cognitive psychology. Computational treatments have concentrated on the conceptualisation (or planning) and formulation (or surface generation) stages, with a bias towards the latter. With the rise of constraint-based representations of grammars and lexica in the 1980s, computational treatments of surface generation have looked towards bidirectional models, in which both parser and generator share the

same linguistic knowledge bases. With a few exceptions however, exemplars have not been built to work within a dialogue context, and little attempt has been made to deal with the reuse of surface structures, or account for incremental production across a number of turns.

A large body of descriptive work now exists for English prosody. Although studies are to some extent diffuse, being hampered by the lack of an appropriate metalanguage for describing both prosodic form and its (largely pragmatic) functions, progress in the experimental field, especially that concerning the prosodic marking of accessibility and turn-taking, serves to constrain possible theories. But whereas a considerable amount of representational work has been achieved at the phonetic level and to a lesser extent the level of syntax and surface prosodic structure, cognitive models of language production are still relatively silent about how pragmatic assignment of prosodic features comes about. The computational treatment of prosody has been largely restricted to assignment of prosodic features to pre-existing texts, as in the systems of Silverman (1987) and Hirschberg (1990). The elaboration of prosody within a model of production is still virtually unexplored.

Chapter 3

Dialogue behaviour and prosody

3.1 Introduction

In this chapter I construct a model of the Agent as engaged in an information dialogue, and demonstrate coverage of a variety of dialogue phenomena. The model favours levels of analysis which are particularly appropriate to the study of intonation, as it affects the dynamics of conversation.

A study of the meaning of contours needs to be set within a framework with respect to which they can be meaningful. The framework is a model of the Agent as information provider, engaged in a dialogue with a Caller. Both need to convey intentions and expectations via the linguistic actions which they produce and receive. In terms of the model, such actions can be thought of as moves which advance its state towards the desired outcome of successful information transfer. These moves, or *dialogue acts* (cf. Section 2.2.1) are represented internally as events containing sufficient contextual information to determine unique transitions between states. At the linguistic level, on the other hand, a speaker uses the resources of language to enable an interlocutor to infer the nature of the intended event, and advance his model accordingly. Thus the prosodic attributes of an utterance may be on a par with the textual attributes, subject to similar Gricean tensions between brevity and informativeness. Prosodic and textual attributes may even be complementary: for example, if a speaker signals an utterance to be a question using inverted syntax, then intonational signals are less crucial; conversely, if the textual extent of the

utterance does not extend sufficiently for this to be done, then intonation becomes an important cue to this function.

The model described here is a rational reconstruction of the Sundial dialogue manager, whose implementation is described in Chapter 5. It is based on an agent's own attempts to impose order on external linguistic events, and respond appropriately.

This chapter begins with an overview of dialogue phenomena, particularly the dynamic and structural aspects, and how they affect speakers' utterances. The study is largely based on analyses of the Flight Enquiries corpus; where appropriate, generalisations about intonational contour are attempted. I then turn to the definition of a computational model of the speaker. Using a symbolic notation to define and describe the behaviour of synchronising concurrent processes, I show how the information provider's behaviour in a variety of dialogue situations may be economically described in terms of the behaviour of a number of computational processes modelling communicating experts or 'agents'. This leads to the definition of dialogue acts, which I take to be the inner characterisation of externally observable linguistic events.

Finally, I present an analysis of contours in the Swedish corpus, and discuss how these relate to a labelling in terms of dialogue acts. The correlations found are of a probabilistic kind; it is therefore concluded that a fully explanatory account would need to consider factors not explored in this thesis.

3.2 Dialogue phenomena

3.2.1 Synchronisation and adjacency

Cooperative conversation needs to be orderly. Turns should succeed one another smoothly, with a minimum of overlap (cf. Section 2.2.1). In addition, cooperative conversation is *structured*; as we have seen in Section 2.2.1, such structure may extend both to exchange-relatedness, and embedding. Evidence for long-term verbatim recall is scant (cf. Section 2.3.2). This might indicate that memory for dialogue

structure is short-lived. Nevertheless, speakers are aware of the unfinished nature of exchanges, of which there may be more than one, over longish stretches of conversation. Example 3.1 illustrates that a response may be delayed for a period of time (2–4: initiative; 13–14 response).

(3.1)

[6] T1:SA:1986 (T)

1	A:	flight information may I help you
2	C:	yes um two eight two bee please
3		um can you tell me if you've got a
4		confirmed arrival time for that
5		(.7)
6	A:	sorry two eight
7		two -was it
8	C:	-((two eight t-wo))
9	A:	-right can you hold on
10		(14)
11	A:	sorry to keep you waiting
12		(.7)
13		two eight two (.3) will be landing now
14		(.) at eleven (.) thirty five

But such long-distance dependency between utterances is unlikely to be purely structural. Consider an example where *A* declines to respond immediately, but offers to ring back, maybe some hours later. *A* may ring up and produce an utterance such as (13–14); it is hard to see however how an argument based on structural coherence can be extended to such a case. Instead, it is sufficient to say that *A* has retained a *commitment to respond* over a period of time. I refer to any pair of moves in which the second discharges a commitment set up by the first, as *initiative-response* (IR-) related. IR-related moves need not even belong to the same dialogue, as we have seen. On the other hand, in Section 3.2.3 I present evidence that responses which are close to their initiatives are more likely to be structurally related to them.

3.2.2 Handling information transactions

From the viewpoint of Caller-Agent interaction, a successful information dialogue is one in which the Caller's goals are satisfied. Assuming that the Agent has privileged access to a special body of information, or *database*, goals may be treated as equivalent to database tasks. An information dialogue consists of one or more cycles

during which a Caller specifies a task, or a task is interactionally specified, and the Agent provides a resolution for that task.

A task may be a query or an update. I concentrate on the former case. Consider for example:

(3.2)
SA4:C1: when is the next flight to rome please
...
SA4:A7: there's a flight this evening at nine

Here the response [SA 4:A7] *directly satisfies* the query [SA 4:C1]. In other words, the information provided by the response serves to instantiate the open proposition which constitutes the query, in such a way that the resulting proposition is true of the database. More complex cases of the query-response relation, than direct (minimal) satisfaction, are *satisfaction of indirect queries*, *near-satisfaction* and *over-satisfaction*. An indirect query is one for which the Agent may need to infer what exactly the Caller needs, for example:

(3.3)
[48] T1:SB:344
8 C: ...the bee ay five
9 eight four from tu^{l}**HL** rin love
...
24 A: it'll be landing hopefully at ten twenty
25 five

To deal with the query of Example 3.3 successfully, the Agent needs to apply the default inference that the Caller is interested in the arrival time of the flight.

Responses that *over-satisfy* requests are common. These provide more information than explicitly requested.

(3.4)
[10] T1:SB:454 (T)
6 C ...I'd like to check on the
7 arrival time of bee ay zero eight four
8 (.) uhm: vancouver seattle (.) to heathrow
...
16 A: yes the flight's on route it's expected now
17 at fifteen fifteen
...
20 terminal four (.3)
21 heathrow airport

A Response may *nearly satisfy* the conditions of the request. Compare (3.3) with the fuller version:

[48] T1:SB:344
8 C: ... the bee ay five
9 eight four from turin love
...
(3.5)
17 A: we've got a five seven
18 ⌈**HLH** nine from turin ⌋ which was scheduled
...
20 A: for **H** ten **HLH** thirty

Example 3.5 is a case where the query as originally specified is unsatisfiable, but a related query (with some of the original constraints relaxed) may be satisfied. In that case a 'cooperative' response (Kaplan 1983, Guyomard and Siroux 1989) is possible. The slightly dispreferred nature of such a response is marked intonationally with **HLH**. The **HLH** accent appears both on the changed element: *five seven nine* and on the part of the utterance which answers the original question: *ten thirty.* This is evidence that the marking is not narrowly associated with the need to modify a constraint, as set-theoretic accounts such as those of Ladd (1980) and Ward and Hirschberg (1985) would suggest (cf. Section 2.5.3); rather the usage seems to be associated with an attitude of deference.

Delayed responses, even if direct, are often marked by some reformulation of the original query. Example 3.1 demonstrates delay occasioned by a confirmation subsequence, and a request to hold the line. Delay may happen for reasons of clarification, confirmation or repair, or simply because of the time taken to look up information in the database. A number of so-called *insertion sequences* (Schegloff 1972) are themselves IR-related, and arise because the task is under-specified:

[10] T1:SB:454 (T)
7 arrival time of bee ay zero eight four
8 (.) uhm: vancouver seattle (.) to heathrow
(3.6)
9 (.)
10 A: **HH** today sir ⌉]
11 (.)
12 C: today

This kind of query (line 10) is unlike the open query, in that a default value (today) is given. I call this a *default query.* Such defaults are easily overturned. It does not seem that a response other than the default is dispreferred, any more than a similar response to an open question would be.

Insertion sequences can be seen to conform to the same pattern as IR-pairs associated with the major task. That is, the notion of satisfaction is central to what constitutes a successful clarification. They are however more likely to form close adjacency pairs.

3.2.3 Interpreting input from the Caller

According to cognitive accounts of discourse coherence (cf. Section 2.3.1) interpretation of utterances takes place against the background of an assumed shared model, which is progressively refined and extended. In Chapter 4 I pursue this further. In addition to discourse-model-related interpretation, cues may be extracted from utterances relating to their function in dialogue. They may contain semantic material of a propositional-attitude nature, as in (3.7–3.8):

(3.7)
[7] T1:SA:2013 (T)
3 C: can you tell me
4 the flight arrival time of bee ay two
5 eight six from (.5) er california (.5)

(3.8)
[28] T1:SB:2082 (T)
22 A: **H** three five $^{\downarrow}$**HL** seven you said
23 C: three five seven yes

Initiatives would appear to be more marked, both semantically and syntactically. *Can you tell me* in (3.7) and *you said* in (3.8) both provide explicit indication of their dialogic function, an open question and a confirmation initiative, respectively. Other cues to initiative status are well-known; these include tags, subject-auxiliary inversion, and the use of interrogative pronouns. There are nonetheless cases where initiatives are not accompanied by explicit cues, and contextual reasoning is needed. In Example 3.9 the major contextual clue that line 7 is an initiative seems to be a negative one: the utterance is not an appropriate response to the only outstanding initiative, a request for information about a flight. Moreover the name of the carrier is something the Caller might be expected to know:

[28] T1:SB:2082 (T)

3 C: ... I wuh- just want to check flights
4 from lyons -coming to
(3.9) 5 A: -yes
6 C: terminal one
7 A: i(k) [↑with british **HH** airways ↑]
8 C: with british airways

In (3.10) the fact that information already given by the Caller is repeated means that A's utterance is probably the initiation of a confirmation sequence:

[34] T2:SA:1045 (T)

(3.10) 7 A: to paris
8 C: yeah

There may however be difficulty in telling whether an elliptical initiative with reduced syntactic features is a default query or a confirmation. The former seem to be marked more consistently with high final boundary tones: cf. (3.6, 3.9), where default queries are all marked with final **HH**, and (3.8), where a confirmation is marked with ↓**HL**. In Section 3.4 I examine more data which points to this distinction.

In the case of responses, surface cues are even rarer. The clearest are discourse markers indicating acceptance or rejection, such as *yes, no, that's right.* These are regularly used to accompany responses to default (or polar) questions, cf. (3.10: l.8; 3.8: l.23). In addition, the discourse marker *yes* is regularly used to mark the accompanying utterance as a response, regardless of the nature of that response:

[37] T2:SA:1235 (T)

(3.11) 16 A: yes I haven't got a flight nine six
17 nine from hamburg ...

In (3.11) the Agent is announcing failure to find a response matching the Caller's constraints. Here and in many other cases, *yes* marks a return to a delayed task, after intervening sub-sequences and holds. See also Example 3.4.

The response *no* is invariably hedged according to some convention of a Gricean nature, whereby if the Caller knows a better response, it would be uncooperative not to give it:

[28] T1:SB:2082 (T)

(3.12) 9 A: ... is it today
10 C: no it's next monday evening ...

Responses which are not explicitly marked as such depend often on adjacency. A further contextual cue may be given by reuse of surface forms:

(3.13)
[5] T1:SA:1356 (T)
44 C: which terminal will I come back
45 A: you'll come
46 to north

Responses are generally intonationally neutral, with falling nuclei. Exceptions are dispreferred responses (cf. 3.5), and responses which are broken across turns, to be discussed in the following section.

3.2.4 Delaying and chunking messages

Dialogue acts, viewed as operators which advance the state of the interaction, are not necessarily co-extensive with turns, nor even sentences.

Firstly, it is possible to have turns consisting of more than one dialogue act:

(3.14)
[13] T1:SB:9588 (T)
61 C: er::m how would I find out will there be
62 any other numbers I can ring
63 A: well that's right I'll just have
64 to give you a general gatwick number
65 just hold on a moment

A's turn (lines 63–65) consists of multiple sentences with related but distinct dialogue functions.

Secondly, a dialogue act may be developed over a number of turns. This tends to happen if a considerable amount of information needs to be transmitted, for example in the case of multiple solutions:

(3.15)
A: there are flights at seven thirty
C: seven thirty
A: eight fifteen
C: eight fifteen
A: and ten twenty

Likewise, information about say an arrival may be too detailed to give in one turn:

(3.16)

[7] T1:SA:2013 (T)

8 A: two eight six from

9 san francisco -is on its way expected

10 C: (-that's th-)

11 A: at **H** thirteen ↑**HLH** ten

12 (.)

13 C: **H** thirteen ↓**HL** ten

14 A: **H** terminal **HLH** four (1) **H** heathrow ↓**HL** airport

It seems that information which the Caller may not be presumed to know is broken down into manageable chunks, the granularity varying with the amount of detail in the information. Thus telephone numbers are typically broken down:

(3.17)

[13] T1:SB:9588 (T)

69 A: ... ring gatwick on **H** oh two nine **LH** three

70 (.)

71 C: oh two nine three

72 A: **H** two **HLH** double eight

73 (.3)

74 C: two double eight

75 (.)

76 A: [↓double **HL** two

This chunking of information for telephone transmission is part of what Clark and Brennan (1991) call *verbatim grounding.*

The data shown here bears out the common observation that when an act is spread over a number of turns, the non-final turns have 'continuation marker' nuclei. These may may be low or high rises or fall rises; a single speaker would seem free to choose which, on a per-turn basis. In (3.17) the Agent uses both **LH** and **HLH** as continuation markers. The end of a broken sentence is frequently marked by a fall to low: cf. Examples 3.16 and 3.17.

3.2.5 Confirmation and repair

Repair moves and repair sequences are occasioned by failure in information transfer. I class with these confirmation sequences, as these presuppose sub-optimal communication. I concentrate on other-initiated repair. Here the party initiating the repair does so as the result of detecting something wrong; otherwise, there is no repair. Assuming that some error occurred in processing the input, there is a scale of possibilities:

1. nothing was understood: the interlocutor needs to request repetition of the entire utterance;

2. part of the input was not understood, but the missing part can be contextualised: this occasions a request for repetition of the missing part;

3. certain elements of the input were doubtful: these need to be confirmed;

4. certain elements of the input correspond to modifications, with respect to the earlier discourse model state: these need to be confirmed;

5. the interpretation of the utterance *qua* dialogue act needs to be confirmed.

I refer to these cases, respectively, as *repetition initiatives, open confirmations, value confirmations, modification confirmations* and *dialogue act confirmations.*

Repetition initiatives frequently use apologetic markers:

(3.18)

[7] T1:SA:2013 (T)
45 A: [↑**H** what ↓**H** saturday is it ↑**HH** for↑]
46 C: **HH** pardon
47 A: [↑**H** what ↓**HL** saturday sir

(3.19)

[26] T1:SB:1772 (T) (P)
6 A: is it **H** heath↑**HLH** row
7 (.3)
8 C: **HH** sorry ↑]
9 A: from **H** heath**HLH** row

Responses to repetition initiatives may reformulate rather than simply echo the previous utterance, as in Examples 3.18 and 3.19. In both cases the textual reformulation is slight. But if the Agent is having difficulty with communication, a more drastic reformulation may be needed.

(3.20)

[42a] T3:SA
4 A: ...is that to**HLH** day madam
5 C: pardon
6 A: to**HLH** day's flight
7 C: it's er yes
8 A: the **H** flight's ↓**H** leaving ↓**HL** today] **LH** yes

The Caller in (3.20) is not a native speaker of English, and communication is on the point of breaking down. The Agent tries a minor reformulation (line 6) which

appears to fail, followed by a major reformulation (line 8), which involves a major change in contour, from fall-rise nucleus, to downstepping to fall. Contour change, or shifting the register up or down, is common in repetitions. In (3.19) there is a nucleus shift on the repeat, and a falling rather than a rising contour. In (3.21) the repeat involves only a nucleus shift. This time though the effect is to emphasize *we* and so permit the implication: *but somebody else may know....*

(3.21)
[39] T2:SA:1931 (T)
17 A: **H** not that we **HLH** know of
18 (.3)
19 C: **LH** sorry
20 A: **H** not that **HLH** we know of ...

The Agent appears to take the repetition request as arising out of a failure of understanding.

Open confirmation initiatives can take a variety of forms, for example:

(3.22)
can you repeat the departure time

flight number what

These share the property of providing enough context for the interlocutor to make out where the gap in knowledge is. Value confirmation initiatives may be echoes with no textual indication of the dialogue act. If however the act is some distance from the earlier utterance which it refers to, explicit marking is more likely:

(3.23)
[2] T1:SA:349 (T) 89 (M)
25 A: ...that's the ell oh two eight
26 one
...
39 C: - what did you say **H** ell oh $\bar{\mathbf{H}}$ tee

(cf. 3.7). Of the unmarked cases of confirmations, not all are responded to. This may be because the speaker forecloses on this possibility by continuing:

(3.24)
[2] T1:SA:349 (T) 89 (M)
43 A: but the flight number is ell oh two
44 eight one
45 (.)
46 C: *sotto voce* two eight (right)
47 *aloud* **H** thanks $^{\uparrow}$**H** ver-y $^{\downarrow}$**H** much in$^{\downarrow}$**HL** deed

Confirmation initiatives between chunks of information (cf. 3.17) are also typically not responded to. Otherwise, hearers do tend to respond to confirmation initiatives, even those for which no explicit marking is available. This suggests that response to confirmation initiatives may be the default; however well marked they are, if they appear in a turn-final position, they get responded to. From the speaker's point of view, failure to use an explicit form such as *did you say ... ?* or intonational marking such as **HLH** may mean that he is indifferent to whether or not he receives a response. If the contextual information which would support an unmarked formulation is not available, then explicit marking will be necessary.

Corrections, and acts requiring that agents update their discourse models non-monotonically, are discussed in Section 4.4.

3.3 A computational model of the speaker in dialogue

In this section I present a reconstruction which reflects the functionality and the architectural principles underlying the Sundial dialogue manager, while differing in details. It also incorporates many of the phenomena observed in Section 3.2. The model is presented as a formal specification; it is not implemented. Its purpose is to bring into a coherent framework a number of the dialogue phenomena observed, and to clarify the notion of 'dialogue act'.

3.3.1 Modelling an information-processing agent

Before considering in more detail internal information processing, I develop a symbolic characterisation of an agent, as it produces and receives messages. The notation of Communicating Sequential Processes (CSP: Hoare 1985) is particularly appropriate for this. An agent can be represented as an autonomous process, whose behaviour is described in terms of the external, observable events which it participates in. An agent which reacts appropriately to the events *hello* and *goodbye* can

be defined algebraically as follows:

$$AGENT \triangleq (?hello \rightarrow !hello \rightarrow AGENT \mid ?goodbye \rightarrow !goodbye \rightarrow SKIP)$$

This definition states that the process *AGENT* reacts to the input *hello* by outputting his own *hello*, or to *goodbye* by outputting *goodbye*. The operator "|" represents deterministic choice between events over which *AGENT* has no control. The definition is recursive, which means that *AGENT* can react indefinitely to inputs of *hello*. On hearing *goodbye*, however, the process *AGENT* becomes the dummy process *SKIP*, representing successful termination.

Two agents modelled in this way will inevitably deadlock, since both will require an utterance from the other in order to do anything at all. This situation can be remedied in the following definition of *DECISIVE_AGENT*, who takes the first initiative.

$$DECISIVE\_AGENT \triangleq (!hello \rightarrow DECISIVE\_AGENT_1)$$

$$DECISIVE\_AGENT_1 \triangleq (?hello \rightarrow DECISIVE\_AGENT_2)$$

$$DECISIVE\_AGENT_2 \triangleq (!goodbye \rightarrow SKIP)$$

Here the process *DECISIVE_AGENT* is described according to the states it passes through: *DECISIVE_AGENT*, $DECISIVE\_AGENT_1$ and $DECISIVE\_AGENT_2$. The conversation between *AGENT* and *DECISIVE_AGENT* can now be modelled in its entirety by the parallel combination of the two processes:

$$CONVERSATION \triangleq (AGENT \parallel DECISIVE\_AGENT)$$

Although concurrent, the processes synchronise on their common events—an output for one agent being an input for the other—so that the following sequence of events can be shown to take place:

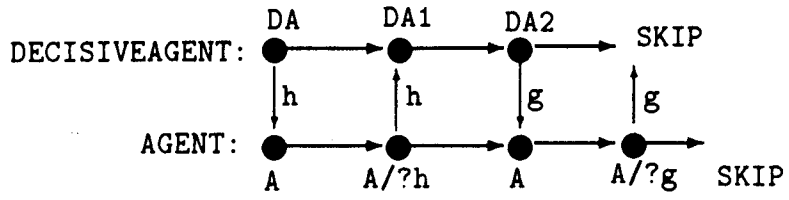


Figure 3.1: State transitions for the AGENT/DECISIVE_AGENT dialogue

(3.25)

Events	Process State changes
$!hello$	$DECISIVE\_AGENT \rightarrow DECISIVE\_AGENT_1$
$?hello, !hello$	$AGENT \rightarrow AGENT$
$?hello$	$DECISIVE\_AGENT_1 \rightarrow DECISIVE\_AGENT_2$
$!goodbye$	$DECISIVE\_AGENT_2 \rightarrow SKIP$
$?goodbye, !goodbye$	$AGENT \rightarrow SKIP$

The derivation in (3.25) represents an account of the conversational sequence: *DA: hello, A:hello, DA:goodbye, A:goodbye.* Events are represented twice, according to whether they are being treated as the output of one process or as the input of the other. A pictorial representation is shown in Figure 3.1. The process *AGENT* is represented as *A*. The notation *P/event* is used to represent the state of process *P* after it has engaged in *event*. *DECISIVE_AGENT* has three states *DA*, DA_1 and DA_2. Thick arrows represent state transitions; time unfolds along the x-axis. Thin vertical arrows are used to represent messages between processes, assumed to pass instantaneously, so that the processes do in fact synchronise on common events. The figure makes clear one property of *DECISIVE_AGENT* that may not be apparent from the definitions: after it has said *goodbye*, it hangs up, and doesn't wait for *AGENT* to respond.

I use the notation of Communicating Processes not to characterise a conversation as a system of communicating individuals, as I have done above, but to describe a single agent—the information provider or Agent in an information dialogue—as a system of internal processes or 'homunculi'. Nevertheless the approach exemplified in (3.25) can be applied in order to achieve internal modelling of the dynamic behaviour of the partner (see Section 3.3.7).

The architecture that I propose is the product of a number of communicating processes each of which are expert in some aspect of conversation or linguistic behaviour. These are the following:

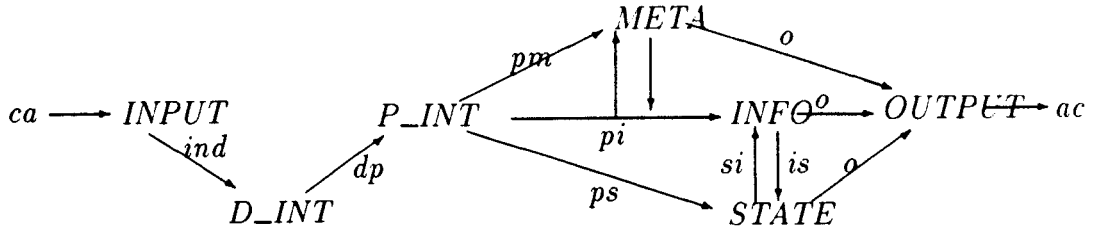


Figure 3.2: Architecture for the Agent as a system of communicating processes

The information component: This is responsible for handling of queries or other tasks of the Caller. It may also initiate its own queries.

The output component: This plans and generates the Agent's messages.

The discourse interpretation component: This interprets the semantic part of the interlocutor's message in the context of the current discourse model.

The pragmatic interpretation component: This uses a combination of cues to assign a dialogue act label—and hence a destination component—to the output of discourse interpretation.

The metacommunication component: This produces repair and confirmation messages as the result of monitoring other messages internal to the Agent; it also handles repair and confirmation messages of the interlocutor.

The complete architecture is shown in Figure 3.2. The acoustic/telephonic channel from Caller to Agent is labelled *ca*; the reverse channel is *ac*. Messages on channels are symbolically described as events annotated by the channel labels; thus $ac.e$ represents the event e on channel *ac*. The notation also allows us to distinguish between $ac!e$ —representing that event from the viewpoint of the outputting process, and $ac?e$ —the same event from the perspective of the receiving process. It is normal to constrain the definition of a process by enumerating the finite set of event symbols, or *alphabet* that it can participate in. I do not attempt to enumerate events for any process except schematically; since they can be characterised as finite feature structures such enumeration can in principle be done. A channel, like a process, has an alphabet: the set of events which appear on it in the definition of

the processes that use it. The alphabets of channels may be assumed to be typed—for example, input to the discourse interpretation component will always take the form of a linguistic sign. For information to be sent along a channel, the message must appear in the alphabet of both receiving and sending processes. This constraint will be exploited in Section 3.3.5, where a process can be inhibited from receiving a message which is flagged as unreliable. Because messages between processes rapidly become complex, these are generally referred to schematically in process definitions, using upper-case variables, hiatuses (...) and indexing. Similar conventions apply in the examples.

The idea of a process communicating along channels can be exemplified by consideration of the *STATE* component. This process is responsible for managing the gross changes of state of the Agent which affect his ability to participate in dialogue, namely: opening and closing the dialogue, and suspending conversation while information is being looked up. *STATE* gets messages from the Caller, after pragmatic interpretation, along the channel *ps*, and sends its messages to the Caller via the channel *o*. A simple definition of *STATE* can be based on that of *DECISIVE_AGENT* above, with *?hello* replaced by *ps?hello* and *!goodbye* by *o!goodbye*, and so forth. For this to have any effect on the remaining components, it must be supposed that as the result of a successful greetings sequence, *STATE* sends out a global message *!begin* to the other components, and that these are all in a quiescent state until having received such a message. Likewise, as the result of the closing sequence, *STATE* sends out the message *!end*, which is included in the definition of every process with the effect that it causes that process to become *SKIP*: ie, terminate. Suspension can be dealt with as follows: when about to do a database access, the component *INFO* sends the message *is!hold*; on receiving this, *STATE* negotiates suspension with the Caller, by sending out the message $o!\langle init;\ hold\rangle$ and awaiting the reply $ps?\langle resp;\ hold\rangle$. Components of the Agent, with the exception of *STATE* and *INFO*, can then be temporarily put into suspension by *STATE* issuing the global message *!hold*. Likewise, once *INFO* is ready to continue, it sends the message *is!resume* to *STATE*, which re-awakens suspended processes with the global message *!resume*, and then negotiates resumption with

the Caller. The events *?hold* and *?resume* must of course be in the alphabets of all relevant processes, with appropriate behaviour defined. Example 3.26 shows typical negotiation sequences:

	A	hold the line please	$o : \langle init; hold \rangle$
	C	thanks	$ps : \langle resp; hold \rangle$
(3.26)	...		
	A	hello	$o : \langle init; resume \rangle$
	C	hello	$ps : \langle resp; resume \rangle$

This also illustrates how a surface form like *hello* can have two related but distinct dialogue functions: beginning a conversation, and resumption after a break.

3.3.2 The Information component

This component (*INFO*) is responsible for communication with the database. It also has the functionalities discussed in Section 3.2.2:

1. Establishing the current task. This may be done on the basis of an initial task formulation by the Caller, or by an explicit request such as *(how) can I help you* or *what information do you require.*

2. Once the task has been adequately specified, retrieving a response, or performing an update.

3. Assessing what constitutes a well-formed query. This means that meta-knowledge about the capabilities of the database needs to be consulted, in order to ensure both efficient look-up and compact response.

4. Sending the well-specified task to the database, and reading back the response. I do not discuss the mechanisms whereby responses other than minimal direct ones are returned, but assume these results are flagged by appropriate features.

5. Accepting further specification, if the initial task formulation is not adequate, or initiating sub-sequences to obtain it.

I discuss first how task information is represented notationally, and hence, what messages (or events) the information component participates in. I then define a

number of relations on task information, before defining the dynamic behaviour of the information component. A query can be notated as follows:

(3.27) $\langle query; Selector, Constraints \rangle$

where *Selector* is used to define the response options, and *Constraints* are used to constrain the query. For example, consider the query corresponding to *what time does sa308 from Rome arrive.* Using *at* to refer to the slot *arrival_time*, *dp* to *departure_place*, *fn* for *flight_number*, this can be represented:

(3.28) $\langle query; \{at : A\}, \{fn : sa308; dp : rome\} \rangle$

Here A represents a named variable whose value is sought; this may be replaced by a value representing a default. Similarly for alternatives queries, the options are specified in the *Selector* field; for example, if asking whether the arrival time was seven fifteen or eight fifteen, the query can be represented:

(3.29) $\langle query; \{at : \{715, 815\}\}, \{fn : ba292\} \rangle \rangle$

The selector therefore both contributes to the constraints, and defines what information is to be extracted as the result of a successful query. One further kind of query is a *polar query*; for this the selector is empty, and the query succeeds or fails depending on whether the information in the *Constraints* field is true or false.

For the two place relation

(3.30) $Satisfies_{DB}(\langle query; \{Sel\}, \{Constraints\} \rangle, Result)$

to be true, the propositions $Sel \cup Constraints$ must be derivable from the database. I do not discuss the presuppositional constraints imposed by the selector. *Result* is a value; for open queries, it represents the value of the unknown variable; for default queries, the value at the path specified by the default selector; for alternatives queries, one of the alternative values proposed. An overloaded response has additionally context-value pairs corresponding to further information which is simultaneously true of the database, together possibly with values for variables which

were left underspecified in *Constraints*, but were not part of the selector. In the case of a polar query, the response (if not overloaded) is empty.

A result to a query can then be divided into three parts, representing the value (if any) for the selector, additional constraints and bindings in the case of overloaded responses, and modifications to the original constraints in the case of approximate responses:

(3.31) $$Result \triangleq \langle Value, Overload, Approximation \rangle$$

Approximate responses may be assumed to contain additional instructions for modifying the original query. In cases where *Overload* and *Approximation* are not present, I show only the result value.

In defining the Information Component, the messages used correspond to events on the channels *pi* (input from Pragmatics Interpretation) and *o* (output to Output Component). Communication with the database is left implicit in the definition of $Satisfies_{DB}$. Possible messages are:

	message	*shorthand*
(3.32)	$pi?\langle init; \langle query; \{Sel\}, \{Cnsts\}\rangle\rangle$	$pi?\langle init^1; Query\rangle$
(3.33)	$pi?\langle resp; \langle query; \{Sel\}, \{Cnsts\}\rangle, \langle result; \{Val\}\rangle\rangle$	$pi?\langle resp^1; Result\rangle$
(3.34)	$o!\langle init; \langle query; \{Sel\}, \{Cnsts\}\rangle\rangle$	$o!\langle init^1; Query\rangle$
(3.35)	$o!\langle resp; \langle query; \{Sel\}, \{Cnsts\}\rangle, \langle result; \{Val\}\rangle\rangle$	$o!\langle resp^1; Result\rangle$

where (3.32) and (3.33) correspond to queries and responses from the Caller, while (3.34) and (3.35) messages initiated by the Agent. Responses thus contain mention of their queries; in the case of responses from Agent, these are needed for proper response formulation; for response from Caller, this is not strictly necessary, so long as queries and responses are indexed as belonging together. One further type of input is simply a set of constraints, without any other information about what can be done with it. *INFO* will try to attach this information to the current task.

I now give a CSP definition of *INFO*. Its states are notated by the current task, ie: $INFO_{\langle Task \rangle}$ where *Task* schematically represents a task specification. If there is no task, the notation $INFO_{\langle\,\rangle}$ is used. I distinguish between states where input is expected, and where output needs to be sent, thus: $INFO_{\langle \ldots \rangle :?}$ for input; $INFO_{\langle \ldots \rangle :!}$ for output. The initial state could either be $INFO_{\langle\,\rangle :?}$, which waits for the Caller to specify a task, or $INFO_{\langle\,\rangle :!}$, which prompts for one. The rules for *INFO* can be stated as follows:

$$INFO_{\langle\,\rangle :!} \;\widehat{=}\; o!\langle preinit;\ Task\rangle \rightarrow INFO_{\langle\,\rangle :?} \tag{3.36}$$

$$INFO_{\langle\,\rangle :?} \;\widehat{=}\; pi?\langle init;\ Task\rangle \rightarrow INFO_{\langle Task\rangle :!} \tag{3.37}$$

$$\begin{aligned} INFO_{\langle Task\rangle :!} \;\widehat{=}\;& \textbf{if } db\_adequate(Task) \textbf{ then} \\ & \quad \textbf{if } Satisfies\_DB(Task, Response) \textbf{ then} \\ & \quad\quad o!\langle resp;\ Task;\ Response\rangle \rightarrow INFO_{\langle\,\rangle :!} \qquad (3.38) \\ & \quad \textbf{else } o!\langle resp;\ Task;\ failed\rangle \rightarrow INFO_{\langle\,\rangle :!} \qquad (3.39) \\ & \quad \textbf{endif} \\ & \textbf{else if } next\_subquery(Task, Subq) \textbf{ then} \\ & \quad o!\langle init;\ query;\ Subq\rangle \rightarrow INFO_{\langle Task\rangle :?} \qquad (3.40) \\ & \textbf{endif} \end{aligned}$$

$$INFO_{\langle Task\rangle :?} \;\widehat{=}\; pi?\langle resp;\ Constraints\rangle \rightarrow INFO_{\langle Task \frown Constraints\rangle :?} \tag{3.41}$$

I define a *pre-initiative* ($\langle preinit; \ldots \rangle$) to be a message which expects an initiative for its response. The unary predicate $db\_adequate(Task)$ is true if *Task* is sufficiently specified for a database access. If this is not the case, the form of an interactive query for further specification is given by the relation *next_subquery*. $Task \frown Constraints$ represents the further specification of *Task* which results from adding *Constraints*.

As an example of the operation of *INFO*, consider the following dialogue:

(3.42)

A1 how can I help you
C1 what time does SA 308 from Rome arrive
A2 is that to Heathrow
C2 no Gatwick
A3 seven fifteen

The events and state changes for *INFO* are as follows:

	Event	*State changes*
*A*1	$o!\langle preinit;\ TASK\rangle$	$INFO_{\langle\,\rangle:!} \rightarrow INFO_{\langle\,\rangle:?}$
*C*1	$pi?\langle init;\ \langle query^1;\ \{at : A\}, \{fn : sa308;\ dp : rome\}\rangle\rangle$	$INFO_{\langle\,\rangle:?} \rightarrow INFO_{\langle\langle query^1\rangle\rangle:!}$
*A*2	$o!\langle init;\ \langle query^2;\ \{ap : hrow\}, \{\ldots\}\rangle\rangle$	$INFO_{\langle\langle query^1\rangle\rangle:!} \rightarrow INFO_{\langle\langle query^1\rangle\rangle:?}$
*C*2	$pi?\langle resp;\ \langle query^2;\ \ldots\rangle, \langle result;\ \{gwck\}\rangle\rangle$	$INFO_{\langle\langle query^1\rangle\rangle:?} \rightarrow INFO_{\langle\langle query^{1\bullet}\rangle\rangle:?}$
*A*3	$o!\langle resp;\ \langle query^1;\ \ldots\rangle, \langle result;\ \{715\}\rangle\rangle$	$INFO_{\langle\langle query^{1\bullet}\rangle\rangle:?} \rightarrow INFO_{\langle\,\rangle:!}$

(3.43)

—where $INFO_{\langle\langle query^{1\bullet}\rangle\rangle:?}$ represents the state of *INFO* when it still needs to answer the Caller's query, but when the task has been refined (by C2). The definition of *INFO* is simple and limited. It assumes that there is only one task at a time. This is not the case, for example, if the Caller asks for some clarification, whilst his main task is still being processed. This situation can be dealt with if *INFO* keeps a stack of tasks, so that with its state represented: $INFO_{\langle Task_1, Task_2, \ldots\rangle}$, $Task_1$ is the current task; once finished it is removed from the stack and $Task_2$ becomes the current task, etc. Also not included are explicit messages for concluding tasks; the Caller might wish to stay with the current task, and ask for more details, after the major response (3.38) has been generated. A more detailed definition would also include the ability to freely mix unsolicited input (including sub-tasks) from the Caller, with requested input.

3.3.3 The Output component

The Output component (*OUTPUT*) is responsible for planning and generating the systems utterances. I have demonstrated (Section 3.2.4) that turns produced by speakers need not be co-extensive with dialogue acts. Since the latter are the main unit of information transfer within the system, this imposes a dual requirement. *OUTPUT* should be able to buffer up several messages before saying any; it should also be capable of breaking up into smaller chunks a single message which contains too much information to be delivered at once. Moreover, if feedback from the interlocutor is to be dealt with in between chunks, *OUTPUT* will be required to suspend its current operation, and make itself available for any repair subsequences

that may occur, for example:

(3.44)
A1: there's a flight this evening at nine
C1: five oclock
A2: no nine oclock
C2: nine oclock
A3: arrive rome eleven thirty
C3: eleven thirty

In (3.44) the Agent's information-providing dialogue act takes place over A1 and A3; C1–C2 is a subsequence involving a repair.

In order to define a mechanism for buffering messages input from elsewhere in the system, I assume a pre-defined limit on the buffer size, *max*, after which the system needs to speak its stored messages. The constraint that speaking is allowed after a fixed number of messages may seem inflexible; compare:

(3.45) A: which airport are you travelling from
what day do you want to leave

(3.46) A: what day do you want to leave
you're travelling from London

(3.47) A: you're travelling from London
what day do you want to leave

Only (3.47) seems acceptable, and that only if the first dialogue act is read as a confirmation not requiring a response. This suggests that speaking should happen as soon as an initiative requiring a response is processed. I shall let $OUTPUT_{n:!}$ be the process after n messages have been stored, and speaking is enabled. The function *generate_messages* takes the buffer contents indexed by its arguments, and produces the corresponding (extended) utterance. The unary predicate *is_init* determines whether or not its argument is a true initiative, ie. requires a response. *OUTPUT* is defined as follows:

(3.48)
$$OUTPUT_{n:!} \triangleq o?message_{n+1} \rightarrow \textbf{if } is\_init(message_{n+1}) \textbf{ then } ac!generate\_messages(\langle 1 \ldots n+1 \rangle) \rightarrow OUTPUT_{0:?} \textbf{ else } OUTPUT_{n+1:!}$$

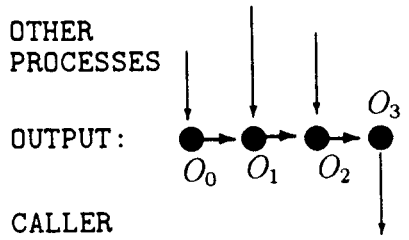


Figure 3.3: *OUTPUT* when the maximum buffer length is 3

endif

$(n < max - 1)$

$$(3.49) \qquad OUTPUT_{max-1:!} \;\widehat{=}\; o?message_{max} \rightarrow ac!generate\_messages(\langle 1 \ldots max \rangle) \rightarrow OUTPUT_{0:?}$$

Buffering is illustrated in Figure 3.3. The process accepts inputs from other processes at states $OUTPUT_{0:!}$, $OUTPUT_{1:!}$ and $OUTPUT_{2:!}$. Its buffer then being full, the three messages are output in a single turn, the process returning to the state $OUTPUT_{0:?}$. Consider now the case where the process is in the state $OUTPUT_{0:!}$, and there are two input messages:

$$message_1 = \langle init; \langle confirm; \{dp : lond\}\rangle; strong : -\rangle$$
$$message_2 = \langle init; \langle query; \{dt : A\}\rangle\rangle$$

The feature *strong* on a confirmation initiative determines whether or not a response is explicitly required (see Section 3.3.8). The state changes of *OUTPUT* are shown in (3.50):

(3.50)

Events	*State changes*
$o?\langle init; \langle confirm; \{dp : lond\}\rangle; strong : -\rangle$	$OUTPUT_{0:!} \rightarrow OUTPUT_{1:!}$
$o?\langle init; \langle query; \{dt : A\}, \{\}\rangle\rangle$	$OUTPUT_{1:!} \rightarrow OUTPUT_{0:?}$
$ac!$"*from london what time ...*"	

The first initiative is not marked [*strong* : +], so the predicate *is_init* is not satisfied, and $message_1$ is retained in an internal buffer. However $message_2$ is an $\langle init; query \ldots\rangle$, which counts as a true initiative, so the composite message: *from London. what time do you want to leave* is generated.

When $OUTPUT_{...!}$ has spoken, the process becomes $OUTPUT_{0:?}$. In this state, further activity is blocked and the interlocutor is free to speak. This may be communicated to the rest of the system via the global message *your_turn*. Conversely we may define an input process, with the two states: $INPUT_?$ and $INPUT_!$. The latter state blocks input from the Caller, because the system is presumed to be speaking; on receipt of the global signal *your_turn* it becomes $INPUT_?$ and is capable of processing Caller utterances. That is:

(3.51) $$INPUT_! \;\widehat{=}\; ?your\_turn \rightarrow INPUT_?$$

I do not enter into details of the process state $INPUT_?$. Its normal activity is to process an acoustic signal from the Caller and return a sign representing the linguistic analysis along the channel *ind* to the discourse interpretation component. On receiving the signal $?my\_turn$, however, or if some period of time elapses during which there is no input from the Caller, it reverts to the blocking state $INPUT_!$. The message *my_turn* may be sent by some component which detects a turn-change signal, or by the interpretation process after analysing an utterance to be a true-initiative.

The message *my_turn* causes a state change in *OUTPUT*:

(3.52) $$OUTPUT_{0:?} \;\widehat{=}\; ?my\_turn \rightarrow OUTPUT_{0:!}$$

As for the phased outputting of chunked portions of a single message, some difficulty can be avoided by the observation that this case should not overlap with that of messages being buffered. I define a process *OUTPUT_CHUNKS* which starts with a sequence of chunks, and whose job is to output these one at a time. *OUTPUT* becomes this new state, if the input message requires chunking; assuming a predicate which decides this and returns n, the number of chunks, the following can be added to the definition of *OUTPUT*:

$$OUTPUT_{0:!} \;\widehat{=}\; o!message_1 \rightarrow$$

(3.53) **if** $requires\_chunking(message_1, n)$ **then** $OUTPUT\_CHUNKS_1^n$ **else** ...

$OUTPUT\_CHUNKS_c^n$ can be defined thus:

(3.54)
$$OUTPUT\_CHUNKS_c^n \triangleq ac!generate\_chunk(c) \rightarrow OUTPUT\_CHUNKS_{c+1}^n \quad (n < c)$$

$$OUTPUT\_CHUNKS_n^n \triangleq ac!generate\_chunk(n) \rightarrow OUTPUT_{0:?}$$

When all the chunks have been generated (using the function *generate_chunk*), *OUTPUT* takes over again.

To allow feedback and possible insertion sequences between chunks, the definition 3.54 must be refined thus:

(3.55)
$$OUTPUT\_CHUNKS_c^n \triangleq ac!generate\_chunk(c) \rightarrow !your\_turn \rightarrow TEMP\_OUTPUT_{0:?} ; OUTPUT\_CHUNKS_{c+1}^n \quad (n < c)$$

where $TEMP\_OUTPUT_{...}$ behaves like $OUTPUT_{...}$, except if it receives a time-out message, in which case it terminates. The infix operator '; ' is used for sequential composition of processes; that is, P_1 ; P_2 is that process which behaves like P_1 until that process terminates successfully, whereafter it behaves like P_2.

Chunking of output is illustrated in Figure 3.4. The process in state $OUTPUT_{0:!}$ receives a message $o?d1$, which requires chunking into three components. The process becomes the process $OUTPUT\_CHUNKS_1^3$ (o_ch_1), which generates the first chunk, sending it on channel *ac* as $m1$. After the global message *yt* (represented as a vertical line), op_1 becomes the process $TEMP\_OUTPUT_{0:?}$, represented as $to_{o:?}$. Were there no input from the Caller, this process would simply die. As it is, the input $ca.s1$ is received, and other processes (not represented) decide to react to it, producing $o.d2$. This is the current active output process, and the

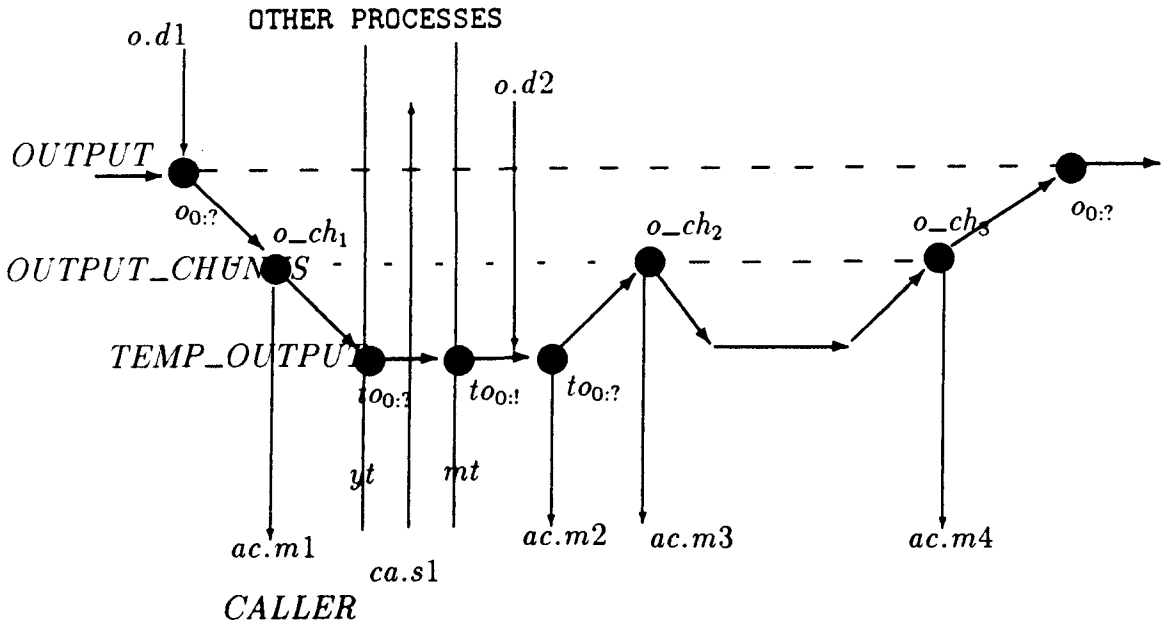


Figure 3.4: Chunking a message

global signal *mt* (my turn) having been received, it is in a state to speak, producing *ac!m2*. *TEMP_OUTPUT* now terminates, and the output process becomes $OUTPUT\_CHUNKS_2^3$ (o_ch_2), which generates the next chunk *ac!m3*. Finally, when the last chunk *ac!m4* has been generated, $OUTPUT\_CHUNKS_3^3$ (o_ch_3) dies, and $OUTPUT_{0:?}$ ($o_{0:?}$) becomes the default output process again.

The account of buffering and chunking offered here is an abstract one, confined to the external behaviour of *OUTPUT*. At which level of language production might buffering or chunking take place? Referring to Levelt's architecture for the speaker (see Figure 2.3), this could be at the level of the Conceptualizer, or that of the Formulator. The definitions given here are sufficiently abstract to leave the issue undecided. In terms of economy of processing, it would seem preferable to make the decision to split output at as high a level as possible.

This section has presented a schematic account of the language production process, from the perspective of its dynamic behaviour within the system. In Chapters 5, I present a detailed account of the workings of the language production model within the Sundial implementation.

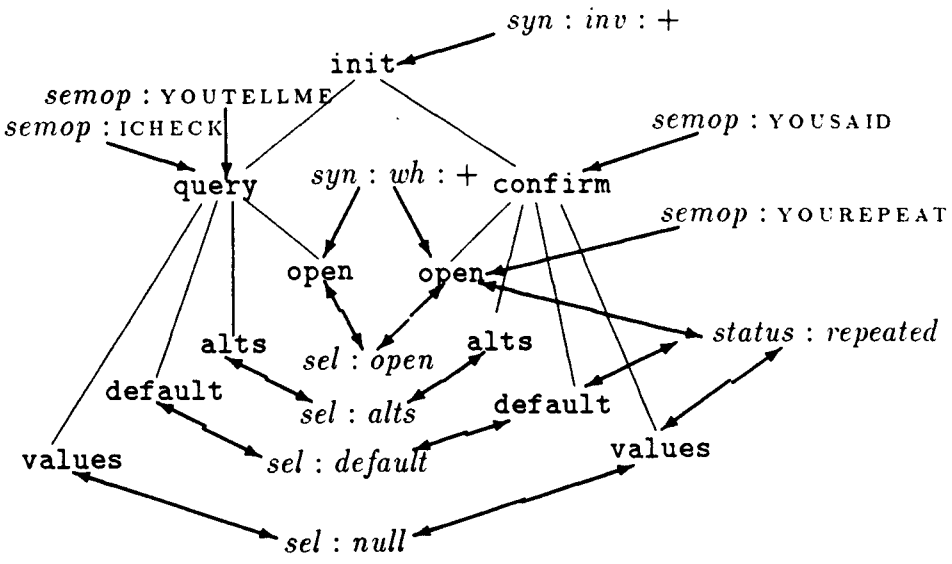


Figure 3.5: Simplified dialogue act hierarchy, showing correlation with syntactic and semantic features

3.3.4 Pragmatic interpretation

The component *D_INT* is responsible for interpretation of surface semantics with respect to the discourse model. I do not present a formal definition of this component; its functionality is discussed in Section 4.3.1.2. The pragmatic interpretation component *P_INT* assigns a dialogue act label to the result of discourse interpretation. As I have shown (cf. Section 3.2.3) the extent to which dialogue function is explicitly marked varies. For responses, it can be marked by *yes/no*; otherwise reliance must be put on matching and adjacency. So far as initiatives are concerned, Figure 3.5 shows some of the correlations between features extracted during syntactic or discourse processing, and dialogue act types. Dialogue acts are organised in a simple inheritance hierarchy, with root `init`. Below this they are divided into queries and confirmations, and these are divided according to whether they are open, alternatives, default or values. Syntactic features are prefaced *syn*. Some semantic features (corresponding in the main to propositional attitude operators) are treated as non-interpretable during discourse interpretation; these are prefaced with *semop*, their values being schematically represented using small capitals. The feature *status* is a flag reflecting the manner in which the discourse model has been advanced: if nothing has been changed, its value is *d_rep*; non-monotonic change

would lead to the value *d_mod*. The feature *sel* refers to selector component of the task constraints derived (cf. Section 3.3.2). Thick arrows denote correlations, unique arrows from a feature node to dialogue act node or *vice versa* indicating that this relation instance is functional. For example, the arrow from $[syn : [inv : +]]$ to `init` indicates that this syntactic feature necessarily implies an initiative. Similarly, the dialogue act $\langle init; query; open \rangle$ necessarily implies the feature $[sel : open]$; the reverse however is not true, because that feature assignment could equally well apply to $\langle init; confirm; open \rangle$.

The mapping shown is complex and far from complete; for example factors like adjacency and the re-use of surface forms have not been taken into account. It nevertheless shows how, even in the absence of matrix phrases or explicit syntactic marking, it is still possible for a hearer to distinguish the dialogue acts at the terminal nodes of this hierarchy, just by the features *sel* and *status*, each of which is available as the result of successful interpretation.

The mapping is also of use to the Agent *qua* speaker; were it not conventional in this way, it would not be possible to exploit it in conversation. In both cases, an algorithm is required which selects the appropriate target nodes: for the speaker, given dialogue act specifications; for the listener, given whatever features have been derived. I therefore assume that the mapping (3.56) is 1–1:

(3.56) $$Surface\_Init : Cues \times Selector \times Status \longleftrightarrow Inits$$

Responses must be dealt with differently. To be counted as such, they must correspond to earlier initiatives. These may be recorded on the *Response Stack*, a datastructure which keeps track of pending responses in a last-in first-out manner, deleting them when they have been found. Items are put on the Response Stack at the time when an initiative is sent to *OUTPUT*. Representing the former by the process *RS* with input channel (from any module) *rsin*, then whenever a response is expected, the corresponding rule must be modified to take this into account:

(3.57) $$\ldots o!\langle init \ldots \rangle \rightarrow Chan?\langle resp \ldots \rangle \rightarrow \ldots$$

(3.58) $\ldots o!\langle init \ldots\rangle \rightarrow rsin!\text{“}Chan?\langle resp \ldots\rangle\text{”} \rightarrow Chan?\langle resp \ldots\rangle \rightarrow \ldots$

That is, if *Chan* is an arbitrary channel, then the expected response in (3.57) must be prefaced in the modified rule (3.58) by a message sending a quoted copy of the template for this response to *RS*. If initiatives are stored in this way, a function $Lookup_{RS}$ within *P_INT* can seek a match between the incoming features and the underspecified template on top of the response stack, and return the resulting instantiated template; this retains the prefix with the address of the component for which it is destined.

P_INT is defined thus:

(3.59) $P\_INT \triangleq dp?\langle Cues, TInf, Status, Validity\rangle \rightarrow$
if $Surface\_Init(Cues, Selector, Status) = \langle query; QSelector\rangle$
then $pi!\langle init; query; QSelector; TInf\rangle \rightarrow P\_INT$
else $Lookup_{RS}(Cues, Selector, Status, TInf) \rightarrow P\_INT$
endif

The message from *D_INT* includes not only a task-oriented interpretation (*TInf*) but also surface cues (*Cues*), status and indication of acoustic validity. The definition is limited in only considering initiatives destined for the component *INFO*; moreover, nothing is said about the possibility of $Lookup_{RS}$ failing to return a result. In that case a failure result may be returned, leading to a request for repetition.

In this section I have attempted to show that the computational task of pragmatic interpretation may depend on a combination of contextual and linguistic cues. This reinforces the suggestion made in Section 3.2.3, that prosodic cues provide complementary information, facilitating pragmatic disambiguation. In Section 3.4 I examine whether good correlations between dialogue acts and prosodic contours exist, which could be exploited by the listener to effect such disambiguation.

3.3.5 The Meta component

Acts that refer to previous acts, or results of processing previous acts, I call *meta-communicative.* In the case of initiatives, these are repetition requests, confirmation

requests (including open confirmations) and corrections. They are treated differently depending on who the initiating party is. Acts initiated by the Caller require the Agent to modify his representations (where necessary), and respond appropriately. Acts initiated by the Agent on the other hand come about because of difficulty arising from processing an utterance of the Caller's. The difficulty needs to be resolved interactively, before processing can continue. I define two processes, $META_1$ and $META_2$ to deal with the two cases.

First, Caller-initiated acts. I assume that the processes *D_INT* and *P_INT* have labelled these successfully. The process $META_1$ then reads in a labelled act on the channel *pm*. The possible labelled dialogue acts are as follows:

$$\langle init;\ repeat\rangle$$
$$\langle init;\ confirm;\ \ldots\rangle$$
$$\langle init;\ correct;\ Mod\rangle$$

In the case of corrections, *Mod* consists of three elements: a context, an old value, and a new value. This needs to be confirmed interactively, before a message to the discourse model *valid_mod* ratifies the change. The definition of $META_1$ is then the following:

$$
\begin{array}{lll}
META_1 & \widehat{=} & (\\
(3.60) & & pm?\langle init;\ repeat\rangle \rightarrow o!\langle last\rangle \rightarrow META_1 \\
 & & | \\
(3.61) & & pm?\langle init;\ confirm;\ \ldots\rangle \rightarrow o!\langle resp;\ confirm;\ \ldots\rangle \rightarrow META_1 \\
 & & | \\
 & & \ldots \\
 & & | \\
 & & pm?\langle init;\ correct(Mod);\ TC\rangle \rightarrow o!\langle init;\ confirm;\ mod(Mod);\ TC\rangle \rightarrow \\
 & & \quad pm?\langle resp;\ confirm;\ mod(Mod);\ TC\rangle \rightarrow md!\langle valid\_mod(Mod)\rangle \rightarrow \\
 & & \quad \rightarrow META_1 \\
 & &)
\end{array}
$$

Here I introduce a further kind of confirmation act, confirmation of a modification. It is characterised by emphatic focus (cf. Section 4.4). Added to the alphabet for *OUTPUT* is the message $o.\langle last\rangle$. This process is required to retain a copy of

the last message generation instruction, at some level during the production process. On receiving $\langle last \rangle$ it formulates this message again; the addition of the feature $[repeated : +]$ however may cause it to be formulated differently.

The operation of $META_1$ may be illustrated by considering the case where the Caller has initiated a correction, with the utterance: *not London Luton.*

Utterance	*Event*	*State changes*
C : *not London Luton*	$pm?\langle init;\ correct(\{\langle dp, lond, lutn\rangle\}$	$META_1 \rightarrow$
A : *from Luton*	$o!\langle init;\ confirm;\ mod(\{\langle dp, lond, lutn\rangle\}$	
C : *yes*	$pm?\langle resp;\ confirm;\ mod(\{\langle dp, lond, lutn\rangle\}$	
	$md!valid\_mod(Mod)$	$META_1$

(3.62)

Here the message $\langle valid\_mod(Mod)\rangle$, where Mod is a modification, needs to be added to the alphabet of D_INT. It acts as a control signal confirming a non-monotonic change.

The definition of $META_2$ is less straightforward. The process needs to monitor for messages where interactive intervention is required, possibly on several channels. I handle this by letting pX be a variable taking as value any one of the several output channels of the process P_INT, including pm. $META_2$ picks up any of these which have been flagged by the earlier interpretation processes as having $[validity : poor]$, interactively confirms them, and puts them back with that feature replaced by $validity : good$. For this to happen, none of the 'consuming' processes of messages from P_INT, including $META$ itself, must allow structures with the feature $[validity : poor]$ in their alphabet. $META_2$ also reads in messages which include the component *failure*, directly on the channel pm; in this case, $AGENT$'s most recent utterance is repeated.

$$
\begin{aligned}
META_2 \ \widehat{=}\ & (\\
& pX?\langle Act[validity : poor]\rangle \rightarrow \\
& o!\langle init;\ confirmact;\ Act\rangle \rightarrow pm?\langle resp;\ confirmact;\ Act\rangle \rightarrow \\
& \quad pX!\langle Act[validity : good]\rangle \rightarrow META_2 \qquad (3.63) \\
& | \\
& pm?\langle failure \ldots\rangle \rightarrow o!\langle last\rangle \rightarrow META_2 \qquad (3.64) \\
&)
\end{aligned}
$$

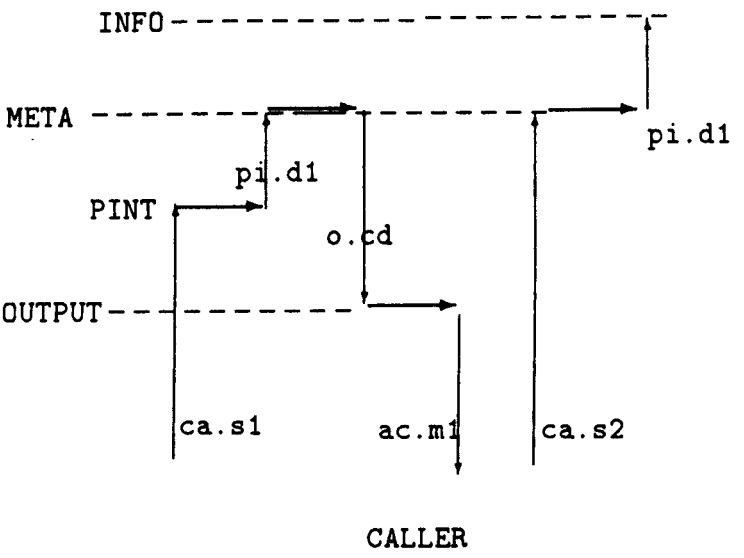


Figure 3.6: The process $META_2$ intercepting a message and requiring confirmation

Here the expression $Act[Feat : Val]$ is the same as Act, except that $Feat$ is required to have value Val.

Figure 3.6 shows the operation of $META_2$, when it interrupts a message intended for *INFO* and requires confirmation. Here and in similar figures arrows are used to represent the flow of control: vertical arrows correspond to messages, horizontal to state transitions within processes. Horizontal dotted lines are used to identify processes. The input message $ca?s1$ is received by *P_INT*; the latter carries out pragmatic interpretation, and decides that the act is a dialogue act $d1$, intended for *INFO*. But the message $pi.d1$ is intercepted by $META_2$, because it carried the feature [*validity* : *poor*]. This then sends out a *confirmact* initiative $o.cd$ to *OUTPUT*, which is in turn generated as $ac.m1$. A *your_turn* signal (not shown) follows, followed by the input $ca.s2$. For the sake of clarity I omit all interpretation stages for this, including pragmatic interpretation, and assume it is sent directly to $META_2$ as an affirmative response. $META_2$ then releases the message $pi.d1$ which it has held back, allowing it to pass to *INFO*, this time with its validity endorsed.

Confirmation of an entire act is the simplest case to handle symbolically; open and value confirmations can be handled similarly. Corrections initiated by the Agent are not included in the definition of $META_2$. To the extent that these are motivated by an interpretation of an unauthorised modification as being a failed confirmation

initiative, this is dealt with in the definition of $META_1$. Otherwise, where the interpretation of some other act is flawed by an unauthorised modification, this may be dealt with by the same kind of interaction that was used to put right acts with poor validity.

3.3.6 An example

The example is taken from the Swedish Airlines corpus. Only processes which change state are shown.

SA4:A1: *Swedish Airlines flight information*

The agent begins as a decisive agent, taking the initiative (cf. Section 3.3.1). *STATE* can be defined to make the response to this $\langle init; opening \rangle$ optional: if the Caller wants to respond with a greeting, he may buffer it together with his first move, C1.

SA4:A1: *can I help you*

Events	*States*
$o.\langle preinit; Task \rangle$	$OUTPUT_{0:!} \rightarrow OUTPUT_{0:?}$
	$INFO_{\langle\,\rangle:!} \rightarrow INFO_{\langle\,\rangle:?}$
	$RS_{\langle\,\rangle} \rightarrow RS_{\langle\langle init;\ Task \rangle\rangle}$

In its initial state, *INFO* has no current task, so issues an $\langle preinit; Task \rangle$ message on channel o. This expects for its response an initiative—shown by the response stack *RS* having on it a template corresponding to this.

SA4:C1: *when is the next flight to Rome please*

Events	*States*
$pi.\langle init; \langle query^1; \{dt : A\}, \{order : next; ap : rome\}\rangle\rangle$	$INFO_{\langle\,\rangle:?} \rightarrow INFO_{\langle\langle query^1\rangle\rangle:!}$

The incoming utterance is analysed as an open query initiative. This matches the expected template on *RS*, which gets popped off.

SA4:A2: *are you travelling from Heathrow*

Events	*States*
$o.\langle init;\ \langle query^2;\ \{dp : hrow\}, \{\ldots\}\rangle\rangle$	$INFO_{\langle\langle query^1\rangle\rangle:!} \rightarrow$ $INFO_{\langle\langle query^1\rangle\rangle:?}$
	$RS_{\langle\rangle} \rightarrow RS_{\langle\langle resp^2;\ Result\rangle\rangle}$
	$OUTPUT_{0:!} \rightarrow OUTPUT_{0:?}$

Within the *INFO* component, the predicate *db_adequate* fails; *next_subquery* produces the default query. Again a template corresponding to the expected reply is saved on the response stack.

SA4:C2: *no Stansted*

Events	*States*
$pi.\langle resp;\ \langle query^2;\ \{dp : A\}, \{\ldots\}\rangle\langle result;\ \{stsd\}\rangle\rangle$	$RS_{\langle\langle resp^2;\ Result\rangle\rangle} \rightarrow RS_{\langle\rangle}$

The response is authorised, and therefore not a correction. The lack of marked emphasis on *Stansted* would probably also lead to this conclusion. The default value in the response template on *RS* is ignored; the result is put on the channel *pi*. However because the default is overridden, it is flagged [*validity* : *poor*], and hence unacceptable in the alphabet of *INFO*.

SA4:A3: *travelling from Stansted*

Events	*States*
$o.\langle init;\ \langle confirm;\ \{\}, \{dp : stsd\}\rangle;\ strong\rangle$	$RS_{\langle\rangle} \rightarrow RS_{\langle pm.\langle resp;\ confirm\ldots\rangle\rangle}$
	$OUTPUT_{0:!} \rightarrow OUTPUT_{0:?}$

The flagged response being in the alphabet of $META_2$, this process initiates a confirmation. Again, expectation of the result is pushed on the response stack.

SA4:C3: *sorry what was that*

SA4:A4: *you're travelling from Stansted*

Events	*States*
$pm.\langle init;\ repeat\rangle$	
$o.\langle last\rangle$	$OUTPUT_{0:!} \rightarrow OUTPUT_{0:?}$

The act is interpreted as a request for repetition; $META_1$ handles this by instructing *OUTPUT* to repeat the last utterance. The presence of the feature [*repeated* : +] forces a reformulation of A3.

SA4:C4: *that's right*

Events	*States*
$pm.\langle resp;\ confirm\ldots\rangle$	$RS_{\langle pm.\langle resp;\ldots\rangle\rangle} \rightarrow RS_{\langle\rangle}$
$pi.\langle resp;\ \langle query^2;\ \{dp : A\}, \{\ldots\}\rangle \langle result;\ \{stsd\}\rangle\rangle$	

The response C4 matches the confirmation response template on the stack; this is then sent to $META_2$, which has delayed sending the response to the default query (C2) to *INFO*. This now goes ahead.

SA4:A5: *hold on please,won't be a moment*

SA4:C5: *thank you*

SA4:A6: *hello*

SA4:C6: *hello*

The hold–resume mechanism was discussed in Section 3.3.1. *INFO* sends out a hold message while a database search is taking place; *STATE* negotiates this with the Caller. Subsequently *INFO* sends out a resume message, and this is negotiated.

SA4:A7: *there's a flight this evening at nine*

Events	*States*
$o.\langle resp;\ \langle query^1\rangle, \langle result;\ \{9\}, \{at : 11\}, \{\}\rangle\rangle$	$OUTPUT_{0:!} \rightarrow$
$ac.generate\_chunk(1)$	$OUTPUT\_CHUNKS_1^2 \rightarrow TEMP\_OUTPUT_{0:?}$

The result of the database search is an overloaded response, which includes an arrival time. In *OUTPUT*, the predicate *requires_chunking* succeeds, and the message is split into two chunks. $OUTPUT\_CHUNKS_1^2$ delivers the first part of the message, then suspends as *TEMP_OUTPUT*, while a response is expected.

SA4:C7: *nine o'clock*

This is treated as an implicit confirmation, not requiring a response.

SA4:A8: *arrive Rome eleven thirty*

Events	*States*
$ac.generate\_chunk(2)$	$OUTPUT\_CHUNKS_2^2 \rightarrow$ $TEMP\_OUTPUT_{0:?}$

TEMP_OUTPUT died because no response was needed to the Caller's confirmation. $OUTPUT\_CHUNKS_2^2$ continues with the second half of the message; again a temporary output process comes into existence.

SA4:C8: *seven thirty*

Events	*States*
$pm.\langle resp; \langle confirm; \{\}, \{at : 730\}\rangle; fail : +\rangle$	

C8 is interpreted in the discourse model as an unauthorised modification; the act is therefore a confirmation which has failed.

SA4:A9: *no eleven thirty*

Events	*States*
$o!\langle correct; \langle at, 730, 1130\rangle\rangle$	$TEMP\_OUTPUT_{0:!} \rightarrow$ $TEMP\_OUTPUT_{0:?}$

The failed confirmation is dealt with by $META_1$, which sends out a correction initiative.

SA4:C9: *eleven thirty*

This is treated as a confirmation of the correction, not requiring a response.

SA4:C9: *thank you very much*

SA4:A10: *thank you*

SA4:C10: *bye*

SA4:C10: *bye*

C begins a pre-closing sequence (C9), which absolves *INFO* from having to send out another $\langle preinit; \ldots\rangle$. The preclosing and closing sequences are handled by *STATE*.

3.3.7 Discussion

The computational system described allows the modelling of a number of properties of human conversation. Firstly, the dialogues that it engages in are structured, with local organisation based on the initiative-response pair, but allowing embedding amongst these. At a global level a dialogue is effectively divided into non-transactional phases: openings, preclosings, holds; and phases during which information transactions can take place. Secondly, the Agent interprets the Caller's goals in terms of what it is capable of; it will prefer approximate solutions to none at all, but will mark these as somewhat dispreferred. Thirdly, because interpretation depends on context, and context is extended dynamically, a set of constraints or requirements can be communicated incrementally, over a number of turns. Lastly, it provides mechanisms for the system to initiate repair in case of failure, and to cope with Caller-initiated repair.

There is in principle no constraint on the degree of embedding of subsequences permitted. This may not be desirable. More than one level of embedding seems particularly unlikely at the purely informational/task level, and has not been observed in the corpora studied. Although the major task may give rise to dependent subtasks, a more complex domain would be required for the dependencies to go any deeper. However if repair moves are taken into account, embedding can go further:

(3.65)

4:C1:	when is the next flight to rome please	*depth* : 0
...		
4:A3:	travelling from stansted	*depth* : 1
4:C3:	sorry what was that	*depth* : 2

The system defined opens itself to a number of interesting extensions. Firstly, although little explicit attempt has been made to model the Caller, a representation is implicitly present, for example in the discourse model, within assumptions about shared knowledge. The architecture presented here also affords the opportunity of modelling an interlocutor's knowledge of his state within the dialogue. For example, a response stack for Caller initiatives can be created as a process that reflects the functionality of the Agent's own response stack. Whenever a Caller initiative has been assigned an appropriate pragmatic interpretation, a placeholder is pushed onto

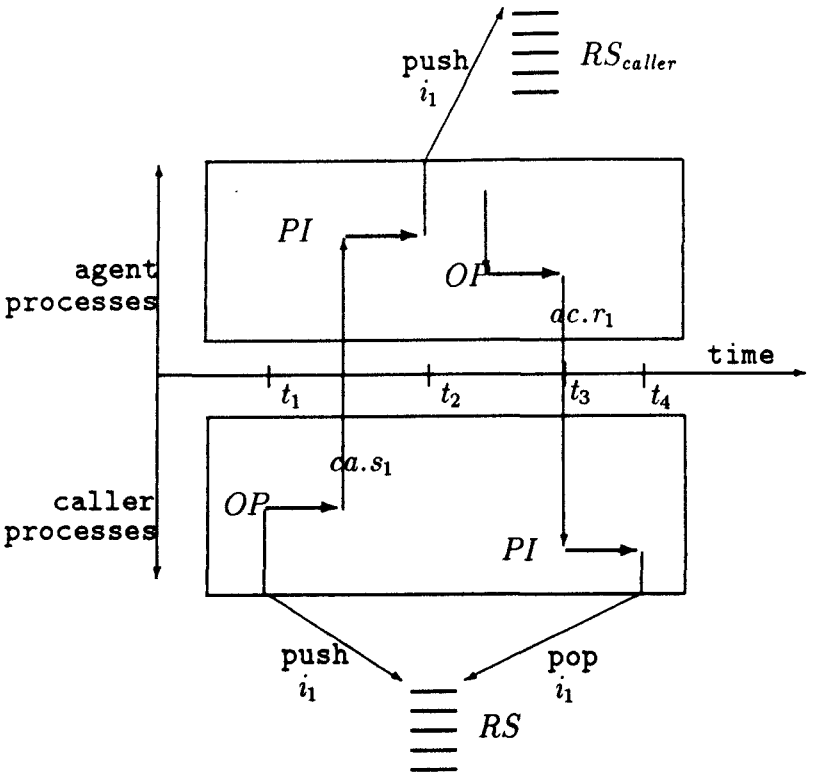


Figure 3.7: Temporal sequence of Caller's result stack and its model by Agent

the Caller response stack; this may be removed once an acceptable response has been provided. This situation is illustrated in Figure 3.7. The Caller's response stack RS and the Agent's model of it RS_{caller} are shown outside the time frame (delimited by rectangles). The figure illustrates the delay inherent in Caller modelling. RS_{caller} only has the message i_1 pushed on at the time t_2, after the Agent has performed pragmatic interpretation of the Caller's message $ca.s_1$. However the Caller updated his own response stack RS at time t_1, when he first formulated the message. Similarly, after pragmatic interpretation of the response $ac.r_1$ at time t_4, the Caller is able to pop the record of his initiative off RS. The issue of when the Agent should do the same with his copy RS_{caller} is not so straightforward. Doing this prematurely (at t_3 for example) entails the risk of the Agent having to revise his model of the Caller subsequently, should the response for some reason be unacceptable to the Caller. A more prudent option would be to delay until after the Caller's next turn, by when it should be apparent if the response has caused any difficulties or not.

The unreliability of conventional phrase-structure representations of dialogue,

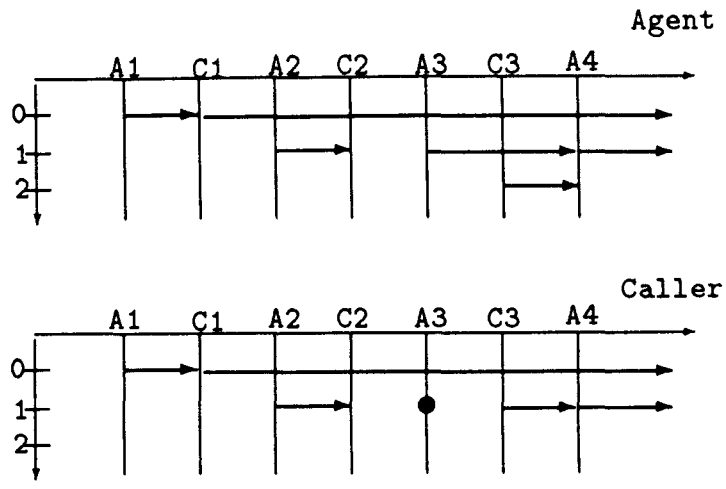


Figure 3.8: Dialogue histories, for both Agent and Caller

as well as the difficulties in modelling the interlocutor, both of which this style of analysis brings out, are illustrated in Figure 3.8, which again refers to Dialogue 4 from the Swedish Corpus. C1 is both a response to an initiative and itself an initiative; similarly for A4, the repeated confirmation request. The relative shallowness of embedding is based on the assumption that the Agent removes the template for the response to A2 from his stack, after receiving C2. If he keeps it there until the response has been ratified, the line A2–C2 needs to be extended leftward, and A3–A4, C3–A4, etc, embedded below it. Also shown in the figure is the structured record that the Caller is able to build up, over the same sequence of turns. It is identical in the initial stages, where the Agent and Caller can be said to agree about what each other's acts are. It differs after A3; the Caller failed to interpret A3 (shown by a dot), and so doesn't need to embed C3-A4 any deeper.

The foregoing discussion has demonstrated that although simple this model may be extended in interesting ways, such as allowing the Agent to maintain a partial and asynchronous representation of some of the interlocutor's internal structures. Some pragmatic inference and revision could then be performed.

3.3.8 A taxonomy of dialogue acts

In terms of the architecture described in this chapter, dialogue acts are best represented at the level of messages emerging after pragmatic interpretation. This is

because pragmatic interpretation is the last link in a chain of refinement which begins with acoustic and linguistic analysis. Similarly on the production side, messages are maximally explicit before being encoded by the output component. Dialogue acts may be defined in two ways: structurally, as bundles of features, or functionally, in terms of the traces they form part of. A *trace* of a process is a sequence of events which the process can participate in (Hoare 1985). I consider first functional definitions of dialogue acts based on traces, before developing a taxonomy along structural lines.

Representing the process $AGENT$ as the concurrent combination of its components as in (3.66), the set of traces $traces(AGENT)$ has as its members all possible sequences of events that $AGENT$ could participate in.

(3.66) $AGENT \triangleq INPUT \parallel D\_INT \parallel P\_INT \parallel OUTPUT \parallel META \parallel \ldots$

An arbitrary trace t from $traces(AGENT)$ contains not only events at the level of dialogue acts, but every kind of message that every component could engage in. A sub-sequence however will consist of dialogue acts. Now for a given dialogue act d_i, a functional definition will relate it to its environment in the traces in which it could have taken part. This can be done either retrospectively, by considering patterns which traces that culminate with d_i should conform with, or prospectively, in terms of patterns on traces which begin with d_i, or in both ways, by considering patterns that constrain traces which include d_i. An example of a retrospective definition would be that a response should culminate a trace which contains in it the corresponding initiative. A response to a correction would be defined in terms of a more constraining pattern, that included a corrective move earlier; to do this the definition of a correction needs to be available. An example of a prospective definition would be that of a pre-initiative. The definition would then constrain traces beginning with such a move to contain an initiative from the other party. Such a definition of course needs refinement, because a speaker cannot guarantee that his interlocutor will understand or respond. This may be achieved by partitioning traces into those that contain preferred outcomes, and those containing dispreferred outcomes.

Corresponding to the functional definition of a dialogue act, a structural definition should specify on it a set of features that can be used to determine whether or not it will meet the prospective or retrospective constraints. It is that combination of features which define the event which the processes that produce or consume it engage in. I do not attempt such rigorous definitions, but define a possible taxonomy of structure-based dialogue act labels. The motivation for this comes from the observations of human dialogues in Section 3.2, and the computational model presented in this section. The basic features are shown in Table 3.1.

Label	Value set	default
owner	self, other	
seqlab	init, resp, preinit, postresp	
type	query, update, confirm, correct, repeat, opening, closing, preclosing, hold, resume	
repeated	+/-	–

Table 3.1: Major dialogue act features

The *owner* feature records which party is speaking. This is important, not least because as we saw in Section 3.3.7, there is a time-lag and a possible lack of symmetry between representations that the two parties may build. The *seqlab* feature takes values corresponding to the sequential position of a move within an exchange: these are *init* (initiative), *resp* (response), *preinit* and *postresp*; the latter representing pre-initiatives (ie, initiatives expecting initiatives) and post-responses (ie, responses to responses). The *type* feature serves to group together dialogue acts, according to the components of the model responsible for handling them:

Component	Types
INFO	query, update
META	confirm, correct, repeat
STATE	opening, closing, preclosing, hold, resume

Finally, the binary-valued feature *repeated* indicates whether or not the dialogue act is being repeated. Default values are provided for some features.

More detailed featural distinctions, which concern the relations between dialogue acts, are shown in Table 3.2 for initiatives, and Table 3.3 for responses. Initiatives that belong to the information component: ie, queries and updates, can be distin-

Type	Label	Value set	default
query, update	topicini	+	
query	qtype	open, alts, default, polar	
confirm	qtype	act, open, alts, default, value	
confirm	failed	+, −	−
confirm	modified	+, −	−
confirm	strong	+, −	+
correct	authorised	+, −	+

Table 3.2: Secondary features: initiatives

Type	Label	Value set	default
any	overloaded	+, −	−
query	relaxed	+, −	−
confirm	confirmed	+, −	+

Table 3.3: Secondary features: responses

guished by whether or not they initiate a task (topic) or are subsidiary. Queries are classed according to what kind of options for response their selector provides; confirmation initiatives likewise. The remaining features: *failed* and *modified* for confirms, and *authorised* for corrects, indicate properties of these dialogue acts with respect to the discourse model. In the case of *failed* and *authorised*, the features are properties assigned by the hearer after pragmatic interpretation, and not present for the speaker. The feature *strong* for confirms indicates whether or not they are marked as requiring a response. As for responses, any of these may be overloaded. It is often possible to represent such overloads as separate dialogue acts. Responses to queries may be relaxed, which is the case when the strict constraints of the query cannot be met, but looser ones succeed. Responses to confirmations may be confirmed or disconfirmed.

Finally, three further features relate to the manner in which dialogue acts are buffered or chunked.

Label	Value set	default
multiturn	+, −	−
dfinal	+, −	+
tfinal	+, −	+

The feature *multiturn* indicates whether or not a dialogue act is spread out over a number of turns. If it is, or if the dialogue act is separated into chunks only by

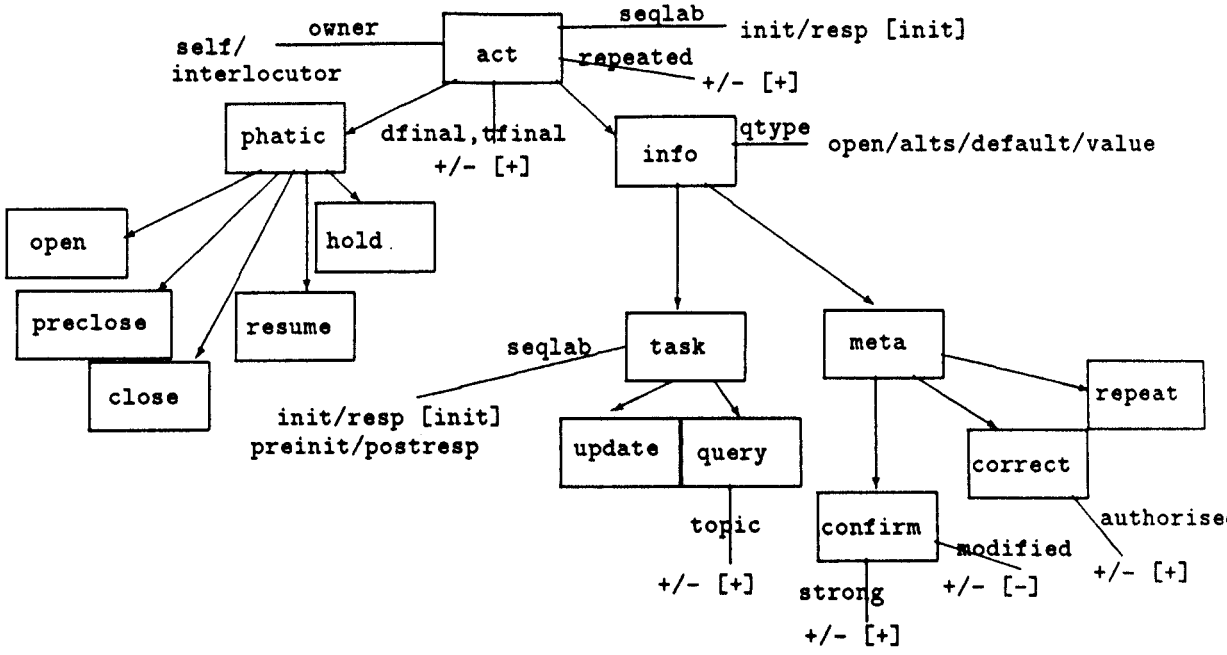


Figure 3.9: Taxonomy of dialogue acts. Defaults in square brackets

pauses, the feature *dfinal* (dialogue-act final) is relevant. If a turn is divided into dialogue acts, then the feature *tfinal* is relevant.

I follow normal practice when defining taxonomies, and make use of inheritance principles to avoid redundancy. The organisation of the taxonomy may be summarised as follows:

1. All acts belong to the inheritance tree. There is one distinguished node, the root *act*, from which all other nodes descend. Otherwise, every node has exactly one parent, and except in the case of leaf nodes, has children.

2. The relation *ISA* between immediately connected nodes is transitive.

3. If A inherits from B, then A inherits all of B's properties, with the addition and possible exception of those more locally defined. Properties or features are atomic-valued. Where possible default values are provided.

Figure 3.9 shows the taxonomy. Arrows represent *ISA* links; simple lines represent featural attachment.

Dialogue act labels should be thought of as bundles of features. Organising them

Updates	[*seqlab* : *init*, *type* : *update*]
Open queries	[*seqlab* : *init*, *type* : *query*, *qtype* : *open*]
Alternatives queries	[*seqlab* : *init*, *type* : *query*, *qtype* : *alts*]
Default queries	[*seqlab* : *init*, *type* : *query*, *qtype* : *default*]
Value Confirmations	[*seqlab* : *init*, *type* : *confirm*, *qtype* : *value*]
Open Confirmations	[*seqlab* : *init*, *type* : *confirm*, *qtype* : *open*]
Corrections	[*seqlab* : *init*, *type* : *correct*]
Repeats	[*seqlab* : *init*, *type* : *repeat*]
Holds and phatics	[*seqlab* : *init*, *type* : *phatic*]
Pre-initiatives	[*seqlab* : *preinit*]
topic major responses	[*seqlab* : *resp*, *topic* : +]
simple Yes/no responses	[*seqlab* : *resp*, *qtype* : *default*]
overloaded yes/no responses	[*seqlab* : *resp*, *qtype* : *default*, *overloaded* : +]
Response to open confirmation	[*seqlab* : *resp*, *type* : *confirm*, *qtype* : *open*]

Table 3.4: Dialogue acts used in the Swedish Corpus

in a hierarchy according to the features *type* and *seqlab* facilitates the representation of a number of regularities and constraints, such as the constraint that phatics are not specified for *qtype*, or that only [*type* : *task*] have *pre* and *post* specified.

Certain properties of the featural system are not so easily represented with an inheritance taxonomy. Thus the feature *seqlab* needs to be specialised for tasks; similar specialisation (not shown here) is needed for the feature *qtype*. Feature cooccurrence relations (Gazdar et al. 1985) are also needed, for example to state that the features described in Table 3.3 are only defined on responses.

3.4 Contours and contexts

This section examines the relationship between the dialogue acts described in the previous section, and intonation contours. The analysis is based on the Swedish Corpus (see Section 1.4.4), which was transcribed both using dialogue act labels, and intonationally. The principle dialogue acts used are summarised in Table 3.4. Intonation is transcribed in terms of sequences of *tonal events* (cf. Section 1.4.2). However, at this level contours are not so easily characterised, or compared. For a broader-based description, a useful starting point is the pitch movement on the 'nucleus' or final sentence accent. A basic opposition between rising and falling nuclei may be observed. Secondary characteristics—such as the direction of the

Downst	Upst	Mixed	Level
22	25	9	41

HLH	**HL**	**LH**	**HH**	Other
61	22	9	3	3

Table 3.5: Open queries

pre-nuclear 'head', and the extent to which individual pitch accents are raised or lowered with respect to one another—may then be compared.

Appendix B gives transcriptions of utterances from the Swedish Airlines corpus, showing variation between speakers. These are grouped broadly according to the taxonomy presented in Section 3.3.8. In the following analysis, I consider first the different kinds of initiatives, as defined in the model, and examine patterns that emerge. Responses are then considered, and compared with initiatives. Finally, conclusions are drawn about the applicability of the analysis to a computer implementation.

Analyses are generally based on all of the transcribed utterances of the corpus. The major exception is that of repeated utterances; these are excluded from other analyses, and considered in a category apart. The featural system does not neatly partition dialogue acts into disjoint categories, and some of the categories considered below overlap. Unless stated, figures given are percentages of all transcribed material of the relevant category, over all speakers.

3.4.1 Task initiatives

Table 3.5 shows the distribution of heads and nuclei for utterances of type ⟨*init*; *query*; *open*⟩. Nuclei are predominantly **HLH**. This trend is reinforced if we disregard speaker JM, who accounts for 71% of **HL** nuclei. The **LH** nuclei all occur with the politeness marker *please*. An alternative analysis for these, with a fall-rise nucleus instead of a fall-plus-rise, would result in an even greater proportion of **HLH** accents. Heads are mainly level; disregarding again speaker JM, who accounts for 43% of all downsteps, the second most frequent category is that of upsteps.

Downst	Upst	Mixed	Level
10	30	10	50

Table 3.6: Head contours for default queries

Downst	Upst	Mixed	Level
42	42	0	14

HLH	**HL**
85	14

Table 3.7: Value queries

Default queries are extremely regular; all speakers use **HLH** nuclei. Table 3.6 shows the distribution of head contours. Value queries (Table 3.7) are mainly **HLH**. There is no clear pattern in the head contours.

Pre-initiatives too have mainly fall-rise nuclei (Table 3.8). There are however a large proportion (16%) of the stylised contour $\mathbf{L}\,\vec{\mathbf{H}}^{\downarrow}\vec{\mathbf{H}}$. These occur with the utterances: [SA 1:A1'; 3:A1']: *can I help you.* A possible explanation is that this act, juxtaposed with a stylised greeting, is normally stylised. Bolinger (1989: 76) refers to this phenomenon, whereby contour spreads beyond the material to which it belongs, as *perseverance* of intonation. Note however that the alternative: [SA 8:A1']: *how can I help you* is only stylised by two speakers.

Alternative initiatives all end with falls. The current analysis ignores act-medial nuclei. Table 3.9 shows heads. All alternative queries (including pre-initiatives) follow the pattern sometimes known as *list intonation*: a number of non-final phrases, followed by continuation markers, then a final phrase, which is generally accompa-

Downst	Upst	Mixed	Level
4	37	12	45

HLH	**HL**	$\mathbf{L}\,\vec{\mathbf{H}}^{\downarrow}\vec{\mathbf{H}}$	$\vec{\mathbf{H}}$	**HH**	Other
58	12	16	4	4	4

Table 3.8: Pre-initiatives

Downst	Upst	Mixed	Level
33	22	44	0

Table 3.9: Alts initiatives

nied by a fall:

(3.67) [SA 6:A8] do you $_1$want to travel $_2$business or $_2$economy class
(JM) $_2$**HLH**] $_3{}^{\uparrow}$**HL**$^{\downarrow}$]

The continuation markers used in the Swedish corpus are fall-rise (**HLH**), low-rise (**LH**) and level ($\vec{\mathbf{H}}$).

With the exception of alternatives queries then, queries tend to have a rising (**HLH**) nucleus. This can be observed in open queries (3.68), default queries (3.69) and value (existential) queries (3.70):

(3.68) [SA 1:C5] $_1$what time does the eight $_2$fifteen arrive
(JQ) $_1$**H** $_2{}^{\uparrow}$**HLH**

(3.69) [SA 6:A2] from $_1$london
(JM) $_1$**HLH**

(3.70) [SA 8:C1] is $_1$there a flight on $_2$sunday $_3$morning $_4$london to $_5$stockholm
(JM) $_1$**H** $_3{}^{\downarrow}$**HLH**] $_4$**H** $_5$**HLH**

3.4.2 Repair utterances

Value confirmation initiatives are divided into strong and weak, depending on whether or not a response is expected. Because of the difficulty in establishing a speaker's intention here, I adopt an operational definition whereby weak initiatives are those which are non-turn-final. Table 3.10 compares confirmation initiatives according to this distinction. Apart from the greater tendency to downstep, there seems little difference between the two categories. The use of $\uparrow$**HL**$_{\downarrow}$**H** would appear to be associated with the violation of a strong default (see Section 4.3.4). In the case of the one speaker (MC) who imposes this interpretation (3.71) the corresponding linguistic event might better be labelled to reflect this.

	Downst	Upst	Mixed	Level
Strong	47	26	0	26
Weak	60	10	10	20

	HLH	**HL**	**↑HL⊥H**	**HH**
Strong	21	73	2	2
Weak	20	80	0	0

Table 3.10: Strong and weak confirmations

Downst	Upst	Mixed	Level
0	57	0	42

HLH	**HH**
85	14

Table 3.11: Open confirmations

(3.71) [SA 3:A2] $_1$five one $_2$two from $_3$paris
(MC) $_1$**H** $_2$**↑HL⊥H**

Open confirmations request repetition of part of a previous utterance. Syntactically, they can be similar to open queries, in using the wh- question form. As Table 3.11 demonstrates, they employ mostly **HLH** nuclei; these tend to be upstepped. The use of high register and upstepping heads means that open confirmations have a lot in common with repeat initiatives. Correction initiatives are distinguished mainly by the preponderance of raised accents; this is in line with the need to make changed material relatively salient, discussed in greater detail in Chapter 4.

3.4.3 Modification utterances

These are marked [*modified* : +], indicating that they refer to some change in the state of the discourse model. In the case of ⟨*init*; *correct*⟩ acts (Table 3.12), the feature *modified* is necessarily present. Table 3.13 shows the distribution of heads and nuclei for all such acts, and for the two sub-cases [*authorised* : +] and [*authorised* : −].

Downst	Upst	Mixed	Level
0	69	7	23

HLH	**HL**
46	53

Table 3.12: Correction initiatives

	Downst	Upst	Mixed	Level
−authorised	0	69	0	30
authorised	11	44	11	33
All	4	59	4	31

	HLH	**HL**	**↑HL⊥H**
−authorised	38	53	7
authorised	11	88	0
All	27	68	4

Table 3.13: Authorised and unauthorised modifications

It might be expected that unauthorised modifications would be generally marked with the 'deferential' fall-rise, as was shown in Example 3.5; this is not the case, although a greater percentage of these have **HLH** nuclei. However, the unit of analysis may again be inappropriate. Nuclei for the utterances

(3.72) [SA 9:A4"] there isn't a flight one five three from geneva
[SA 9:A4"'] there's a flight one nine three

are as follows, for the four speakers:

	9A4"	9:A4"'
JM	**HLH**	**HL**
JQ	**HLH**	**HLH**
MC	**HL**	**HLH**
MG	**HLH**	**HL**

It can be seen that every speaker uses a **HLH** nucleus at least once in a pair. On the other hand, for [SA 8:A7]:

(3.73) well, there's a flight at eight fifteen,

Downst	Upst	Mixed	Level
18	33	7	40

HLH	**HL**
22	77

Table 3.14: Heads and nuclei for repeated utterances

	HLH→HL	**HL→HLH**	**HL↓]→HL**	X→X	⇑	⇓	+phrase	nochange
Inits	10	10	10	70	55	10	20	10
Resps	0	0	0	100	0	50	0	50
All	9	9	9	72	50	13	18	13

Table 3.15: Repeated utterances compared with antecedents

only one speaker (MC) used an **HLH** contour. Further evidence concerning the use of **HLH** accents in these dialogues comes from a study by House *et. al.* (1992). In an experiment designed to evaluate the importance of pragmatically appropriate intonation in speech synthesis, listeners were presented with short dialogue excerpts based on exchanges taken from the Swedish Corpus. In the case of sentences [SA 9:A4"], [SA 9:A4"'] and [SA 8:A7], there was a significant preference for material synthesized with the **HLH** accent, over tokens using the default fall.

3.4.4 Repetitions and reformulations

Table 3.14 shows the patterns of heads and nuclei for repeated utterances. It may be more informative to consider the changes between repeats, and their antecedents. Table 3.15 shows contour changes, whether accents have been boosted (⇑) or depressed (⇓), and whether the number of phrases has increased (+phrase). There is also a category for no change. For each category, percentages of the relevant total are given. Contour changes seem to be confined to initiatives. Changes such as an increase in the number of accents were not taken into account; in the corpus these were also associated with a change in the textual form of the utterance. The utterances for which no (prosodic) change was noted: [SA 2:A4;7:A3] were in fact textually different from their antecedents. Discounting these, it is possible to con-

	Downst	Upst	Mixed	Level
Inits	33	27	8	30
Resps	29	13	6	50

	HLH	**HL**	**LH**	**↑HL⊥H**	**L$\bar{H}$↓$\bar{H}$**	**$\bar{H}$**	**HH**	Other
Inits	38	44	7	0	2	1	2	2
Resps	8	89	0	0	0	1	0	0

Table 3.16: All initiatives and responses, excluding repeated utterances

clude that in reformulations and repetitions, prosodic change is a complementary cue to textual change.

The data of Table 3.15 only compares nuclear tones which are either act-final or turn-final. Thus alternative questions, which always end with **HL**, are classed as X→X. Such analysis masks the fact that these tones tend to be boosted in the repeat. It also fails to take into account the continuation tones before the final fall, for example, $_2$**HLH** in (3.67). Of these, 4 out of 6 show a change in contour, the most common being $\bar{\mathbf{H}}$**→HLH**.

3.4.5 Responses

Table 3.16 compares heads and nuclei for all initiatives and responses, except repeated utterances which were again excluded from the analysis.

The pattern that emerges is one where responses are predominantly marked with falls, while initiatives are fairly evenly divided between **HLH** and **HL**. Given the lack of structural variation among responses, we may take them to be the unmarked form. This is in accordance with the account of the contextual interpretation of responses (Section 3.3.4); compared with the interpretation of initiatives, this is relatively straightforward.

Dialogue acts in which a lot of information is communicated may be broken up over a number of turns, as we have seen (Section 3.2.4). In the Flight Enquiries corpus, [*dfinal* : −] turns were frequently marked with continuation tones, such as **HLH**. In the Swedish Corpus however, both [*dfinal* : −] and [*dfinal* : +] turns end in falls. Differences are subtle: the final turn in a chunked sequence is more likely

	Downst	Upst	Mixed	Level
dfinal	71	14	14	0
$-$*dfinal*	0	37	12	50

Table 3.17: Distribution of heads according to act-finality

to have a nucleus that is downstepped or falls to the bottom of the speaker's range. For one speaker, the [*dfinal* : $-$] falls were less steep than the final ones. Table 3.17 shows the distribution of heads. Non final turns tend to be upstepped or level, whereas final turns tend to be downstepped; in these cases heads may be providing the cues not provided by nuclei.

3.4.6 Discussion

The above analysis reveals little in the way of reliable correlations between dialogue act and contour. There are interesting regularities, but none that could claim an explanation within the model. Consider the open query [SA 5:A2']

(3.74) $_1$what time $_2$do you want to $_3$leave
(JQ) $_1$**H** $_3$**HLH**
(MC) $_1$**H** $_3$**HL**

What distinguishes the tokens uttered by JQ and MC can hardly result from any distinctions made in the model presented in Section 3.3. An attempt at explanation would need to be based on concepts such as attitude, or interpersonal relations; these are by definition excluded from the idealised notion of information dialogues.

The analysis thus vindicates the stance taken by Cutler (1977), Cutler and Isard (1980), Couper-Kuhlen (1986) and Bolinger (see for example Bolinger 1989: 2), that intonation contours are not amenable to analysis along the lines of intentions, or speech acts. Even phenomena which previous research would predict to be present, such as dialogue-act-medial signals, and the use of the fall-rise to mark cooperative responses, were not found consistently.

Such equivocal results may nevertheless be of use in an implemented speech output system, where contour needs to be assigned, even if heuristically. In the Sundial implementation, the contour assignment rules, described in Section 5.6, are

based on a single speaker. Averaging over speakers would not achieve any generality, and would be likely to blur within-speaker distinctions. Data not considered in detail here, including the use of stylised contours to accompany phatic utterances (House and Youd 1991, House 1992) have been added. An attempt has been made to exploit the opportunities for representational economy offered by the inheritance hierarchy. For example, default rules appear for initiatives and responses which are not further specified, or whose specification fails to match higher-ordered rules. Variability in the rules (reflecting the percentages observed) is modelled by weighted random selection. The implementation therefore, whilst lacking explanatory power, is nevertheless consistent with observations.

3.5 Summary and Conclusion

This chapter first examined dialogue phenomena from a structural and prosodic point of view, with reference to the flight enquiries corpus. This led to a specification for a computational model of an information-providing Agent in dialogue. The model can be viewed as an idealisation of the Sundial manager. In Section 5.1.3 the two are compared. The model can be applied in the analysis of the constructed dialogues of Appendix A; an example is given in Section 3.3.6. As a result of the specification, a featural notation for dialogue acts was proposed, with certain features corresponding to events in the model. Features relating to high-level functions (such as the default nature of a query) may be combined, in the description of dialogue acts, with features relating to lower level formulation decisions (such as non-finality). The Swedish Dialogues were labelled according to this notation, and transcribed intonationally, to determine the extent to which regularities between dialogue act descriptions and contours existed. Although some regularities were found, these were lacking in predictive power.

It does not therefore seem possible to account for contour choices in any determinate way, in terms of the model. This negative result should however be seen in the following light: the analysis has simply shown that the phenomena are more complex than a model such as the current one can handle. If, as a number of researchers

have suggested, a speaker's attitude is a central factor, then an exploratory account must await the successful modelling of attitude. Such an exploration has, however, value in offering rules, albeit probabilistic, that will generate contour choices, given an abstract characterisation of dialogue moves (see Section 5.6). This itself is an advance on the bland heuristics generally used in text-to-speech prosody generation.

Chapter 4

Focus assignment

4.1 Introduction

In this chapter I explore how the decision to make items within an utterance more or less prosodically prominent may come about, on the basis of internal contextual representations which are independently motivated, to a greater or lesser extent, by considerations of the computational task faced by the speaker during the process of language production. In the previous chapter although an explicit model of dialogue was proposed, the analysis of prosodic contour was inconclusive. By contrast, prosodic prominence is better understood (cf. Sections 2.5.1–2.5.3). I therefore elaborate a model which is more theoretical, and which attempts to account for prosodic prominence as an emergent property of the whole process of utterance production.

Attempts to provide structural accounts of the incidence of prosodic accent have been of limited success. Even semi-structural accounts such as that of Gussenhoven (1984) are so hedged with exceptions as to carry little explanatory power. I follow Bolinger (eg. 1989) in preferring non-structural accounts of prominence, and relative prominence, for structures within which distinctions may be made at at a conceptual level, or in terms of 'interest'. This is possible because the representations I present cover in some detail the conceptual level, and the level of conceptual-lexical correspondence. Because of this, there is no need for structural rules of projection or percolation.

Focus assignment may come about in a number of ways. Firstly, elements may

be assigned focus properties according to the re-use of previous linguistic forms in their formulation:

(4.1) When do you want to leave
...
When do you want to ≪arrive≫

(1.7') ...leaves at seven twenty in the evening our time
>which is< ≪eight≫<twenty> >in the evening< ≪their≫>time<

The assignment of relative prominence can be seen as a signal to the hearer indicating how the current phrase was built up out of past productions, and so assisting him in rebuilding the structure during parsing. Evidence discussed in Section 2.3.2 supports the notion that previous surface forms may be temporarily retained, and used for this purpose. In Section 4.2 I propose how surface generation mechanisms may take advantage of this, and how prosodic focus assignment in turn can be based on it.

Secondly, elements may obtain prosodic marking according to their *accessibility in the discourse model* (cf. discussion in Section 2.3.1). Broadly speaking, this means that conceptual entities at the level of the discourse model may be ordered for focal prominence: the more accessible, the less prominent. For example, in this exchange:

(4.2) *C*: i want to book a flight to Heraklion
S: do you want to fly business or economy class to >Heraklion<

the discourse entity corresponding to *Heraklion* has recently been mentioned, is therefore highly accessible and so gets defocussed. That we are dealing here with accessibility at a conceptual level as much as at a surface level, may be illustrated if we replace the token of *Heraklion* in *C*'s utterance with *Crete*. In a discourse context where both speakers have the world knowledge whereby so far as airports are concerned, Crete and Heraklion are equivalent, the defocussing would still go ahead. In Section 4.3 I propose a representation of the discourse model in which discourse entities are represented detached from surface linguistic features, and so as to persist throughout the discourse. The relative accessibility of entities in the discourse

model, together with conflicts between information states, leads to a definition of a prominence ordering over discourse entities.

Thirdly, the decision to mark certain message elements for focus may form part of message genesis—ie, be a direct result of the decision to speak. For example, a *correction* will place contrastive focus on the element requiring modification:

(4.3) The flight arrives ≪at nine≫
(not at five)

Similarly, for a *confirmation* request those message elements to be confirmed are required to be prominent:

(4.4) You want to fly <from london> <on sunday>

Section 4.4 takes into account the prominence needs of discourse elements which are central to the speaker's intention.

These levels appear to correspond to the three levels of discourse structure discussed by Grosz and Sidner (1986). At the *intentional* level, an utterance is planned as a means towards solving some goal of the system. This goal may require singling out certain message elements for special attention; for example, a request for confirmation requires that the elements in need of confirmation are marked specially. At the *attentional* level, as descriptions of message elements are planned with reference to the context represented in the discourse model, concepts are ranked for prosodic prominence on the basis of their accessibility in the discourse model. A concept which is highly accessible will thus be relatively low in the prominence order. At the *linguistic*, or *textual* level, generation of the surface structure of an utterance may be facilitated by referring to a previous utterance with which it shares structural and semantic features. If this is the case, certain constituents of the utterance will be assigned prominence according to the manner in which they re-use, or modify, the previous material.

According to most accepted models of production, the intentional, attentional and textual/linguistic levels are processed in that order; in this discussion however, I shall begin with the textual level, moving on to the attentional and intentional

levels. This order of presentation will facilitate the introduction of representations and mechanisms.

This chapter presents a cognitive account of the production of prosodic focus, viewed as part of the larger process of language production. Apart from Section 4.3.2, where the effects of recency and inferrability in the Swedish Airlines corpus are considered, the analysis of data is informal. I assume phenomena of focus and accent to be largely independent of those issues of contour choice examined in Chapter 3. Only cases of local contour choice are considered, where it may be used to distinctively mark simple prominence information. No attempt is made to account for the variability found in the corpus. In Chapter 5 I describe an implementation in which many of the more theoretical points made in this chapter are realised. Specific claims and ideas presented here are, wherever possible, accompanied by references to details of that implementation.

4.2 Focus assignment by reference to surface forms

4.2.1 Modelling the retention of surface forms

We have seen (Section 2.3.2) that speakers are attuned to surface features of their own or their interlocutors' previous productions; but that this retention is short-lived, unless it serves a purpose. A simple model of recall uses a buffer of previous utterances, produced by either party. We can limit accessibility (and possibly the length of the buffer) to the few most recent clauses. I shall refer to this datastructure as the *linguistic history.*

In a model of an Agent who is alternately speaker and hearer, the linguistic history can be of benefit in either modality. If a previous encoding of information is shared and persists, then there is reduced effort for the speaker in reusing that encoding, and correspondingly reduced effort for the hearer. In Section 4.2.4 I shall sketch how a speaker and a listener might reuse previous expressions; first of all though, it is useful to consider the nature of the dialogue contexts in my material, in which examples of reuse have been found.

A speaker may echo back information just received:

(4.5)
[7] T1:SA:2013 (T)
A: two eight six will be landing now at thirteen ten
C: thirteen ten
A: yes

In (4.5), the echoing utterance may have been intended as a request for confirmation; it was probably taken as such by the Agent. In contrast to face-to-face conversation, confirmatory echoing is very common in telephone dialogues. It may also be used in a restatement of information after a confirmation request:

(4.6)
[3] T1:SA:632 (T)
A: yes it'll be landing now ahead of schedule ...
C: ahead of schedule bee ay two nine six
A: ahead of schedule yeah

A related use of echoing, less likely to initiate interaction, is the *back-channel* signal. In the following example, the Agent overlaps the Caller:

(4.7)
[2] T1:SA:349 (T) 89 (M)
A: in future you need to dial seven five nine one eight one eight
C: one eight one eight
A: now let me just look for you

Such uses of echoing are common in telephone dialogues, where they may indicate correct transmission of information.

The examples of echoing discussed so far occur across speakers. Self-repetition, on the other hand, is sometimes necessary when a speaker hasn't succeeded in obtaining the intended response or attention from the interlocutor:

(4.8)
[2] T1:SA:349 (T) 89 (M)
C: ... right thanks very much indeed
A: now when you've got to ring tomorrow you ring 759 18 18
C: right okay thanks very much indeed

In (4.8), the Caller appears to have the goal of terminating the conversation. The Agent isn't quite ready, which accounts for the Caller's need to repeat the pre-closing move. Self repetition occasioned by a lack of appropriate response to an earlier dialogue act may result in reformulation of intonation, as was discussed in Section 3.4.

Echoes and self-repetitions explicitly reference earlier utterances. The same may be said of *substitution* utterances; however in these, whereas a portion of the antecedent expression may be retained, one or more subexpressions are changed:

[31] T2:SA:328 (T)
(4.9) *A*: ...it is scheduled for thirteen thirty
C: yeah I think it was scheduled to fly at twelve fifty

Utterance pairs such as this can be distinguished from echoes, not so much because the members of a pair do not have identical wording, as because they convey different information, and this difference may be marked. The substitution may occur within a single turn; especially in cases where parallel structures are being used:

[2a] T1:SA: (T)
(4.10) ...either the four seven three which is scheduled for seventeen thirty
or the four seven five which is scheduled at ten o'clock

4.2.2 Representing linguistic and propositional information

Before examining in further detail the mechanisms that might be used by speakers and hearers in the re-use of previous surface forms, I present a theory of linguistic representation capable of expressing the necessary multilevel constraints between lexis, syntax, and semantics.

This notation, based on the principles of *unification of partial information*, and *lexicality*, has been influential in computational linguistics since the late 1970's (Kay 1984, Gazdar et al. 1985, Bresnan and Kaplan 1982). The version described here owes most to Head Driven Phrase Structure Grammar (HPSG: Pollard and Sag 1989) and Unification Categorial Grammar (UCG: Calder et al. 1988b). The basic unit of encoding, the *sign*, is divided into fields representing various levels of linguistic knowledge:

$$sign \equiv \begin{bmatrix} phonology \\ syntax \\ semantics \end{bmatrix}$$

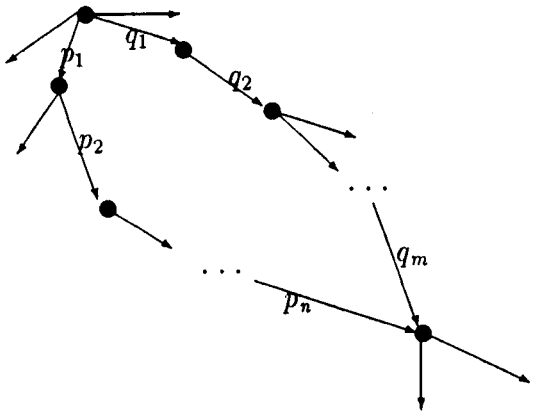


Figure 4.1: Reentrancy in a graph: paths $\langle p_1, p_2, \ldots p_n \rangle$ and $\langle q_1, q_2, \ldots q_m \rangle$ point to the same node

The **syntax** component is recursively defined to contain signs, representing constraints on the environment of the word or phrase represented by the sign. Constraints between components and levels, enforcing agreement, dependencies, coindexing, etc, are implemented in the representation via the use of *re-entrant feature descriptions*, whereby distinct feature-labelled paths point to the same value, as exemplified in Figure 4.1.

A sign in general is therefore not a feature tree, but a *feature graph* (Shieber 1986); unlike trees, in which nodes can have at most one parent, the more general category of graphs allows nodes with several parents. Utterances have syntactic and semantic structure by virtue of lexical entries being encoded as signs which may be complex. A complex sign, or *functor category*, both constrains its environment by specifying syntactic and semantic patterns that its subcategorisation frame must satisfy, as well as describing how its semantic field is built as a function of these components. A simple example is the entry for *ninety*, to be used as functor in structures such as *ninety seven* in which a unit argument follows.

$$\begin{bmatrix} \mathit{phon} : \mathit{ninety} \\ \mathit{syn} : \begin{bmatrix} \mathit{head} : \mathit{noun} \\ \mathit{args} : \left\langle \begin{bmatrix} \mathit{syn} : \mathit{head} : \mathit{noun} \\ \mathit{sem} : S \\ \mathit{dir} : \mathit{post} \end{bmatrix} \right\rangle \end{bmatrix} \\ \mathit{sem} : \begin{bmatrix} \mathit{id} : \iota \\ \mathit{type} : \mathit{tens} \\ \mathit{value} : 9 \\ \mathit{next} : S \end{bmatrix} \end{bmatrix}$$

This states that the functor is followed by a single argument with semantics S, and this semantics is copied into the $<$ *next* $>$ field of the resulting semantics.

If required, the argument can be further constrained, say to having semantic type *units*.

The constraining and structure-building aspects are both described locally, in the lexical entry. There is therefore no need for a separate set of phrase structure rules; instead these are replaced by highly generic rules of combination for signs, which are based on categorial grammar rules of combination: in particular, backward and forward functional application:

(4.11) $$A/B\ B \rightarrow A$$

(4.12) $$B\ A\backslash B \rightarrow A$$

The operator symbols '/' and '\' indicate order of application; the notation can be streamlined somewhat by defining

$$A/B_{prec} \triangleq A\backslash B$$
$$A/B_{post} \triangleq A/B$$

That is, the directionality of the 'slash' operator is specified by a feature on the argument. When the order of an argument B with respect to its functor A is unknown or unimportant, the functor argument relationship can be indicated simply as A/B. As a further notational device, $((\ldots(A_0/A_1)\ldots)/A_{n-1})/A_n$ can be replaced with $A_0/\langle A_n, A_{n-1}\ldots A_1\rangle$, or in a feature structure encoding:

$$\begin{bmatrix} head : A \\ args : \langle A_n, A_{n-1}\ldots A_1\rangle \end{bmatrix}$$

The reverse stacking of the arguments is common in sign-based approaches, and is used as the default order of reduction of the sign with respect to its arguments.

The atomic version of functional application may be extended to complex feature structures, as in UCG:

$$(4.13)\quad \begin{bmatrix} phon : P \\ syn : \begin{bmatrix} head : H \\ args : Args \end{bmatrix} \\ sem : S \end{bmatrix} \quad Arg \quad \rightarrow \quad \begin{bmatrix} phon : APPLY[P, Arg.phon] \\ syn : \begin{bmatrix} head : H \circ \theta \\ args : \overline{Arg_i} \circ \theta \end{bmatrix} \\ sem : S \circ \theta \end{bmatrix}$$

$$where \quad Arg_i \in Args \wedge$$
$$\theta = UNIFY[Arg, Arg_i] \wedge$$
$$\overline{Arg_i} = Args - Arg_i$$

In other words, the functor sign can be reduced with respect to the argument *Arg* by unifying *Arg* with one of the arguments *Args*, discarding that argument, and applying the resulting substitution θ to everything that remains.

For example, consider the phrase:

(4.14) Air France operates from Stansted

For the sake of the present discussion, I treat the phrase "Air France" as having a single atomic lexical entry. I use the shorthand in (4.16) to refer to a structure schematically given as (4.15), where Args is a list of signs, and Args1 the corresponding shorthand expression. When the phonology is underspecified, the shorthand form is as given in (4.17); where *ARGS* is empty it is as given in (4.18), and so forth.

$$(4.15)\quad \begin{bmatrix} phon : P \\ syn : \begin{bmatrix} head : MAJOR : Cat \\ args : Args \end{bmatrix} \\ sem : S \end{bmatrix}$$

(4.16) P : **Cat**/*Args*1 : s

(4.17) **Cat**/*Args*1 : s

(4.18) P : **Cat** : s

The required lexical entries for (4.14) are then the following:

(4.19) *Air France*:**np**:AIR_FRANCE
operates:**s** / [**np**:CARRIER, **pp**:AIRPORT]:OPERATION(CARRIER,AIRPORT)
from:**pp** / [**np**:S]:S
Stansted:**np**:STANSTED

The derivation may be described as follows:

(4.20) *Air France operates from Stansted*: s:OPERATION(AIR_FRANCE,STANSTED)
operates: s / [np:CARRIER, pp:AIRPORT]:OPERATION(CARRIER,AIRPORT)
Air France: np:AIR_FRANCE
from Stansted: pp:STANSTED
from: pp / [np:S]:S
Stansted: np:STANSTED

A theory of linguistic representation based on the sign is appropriate to describing the relations between linguistic and semantic information, for a number of reasons:

1. it is declarative, and therefore potentially bidirectional. It may therefore be used to describe the competence of both a language analyser and a language encoder, on the assumption that these ought to be identical;

2. words and phrases may be simultaneously described at a number of levels. This will enable us to relate intonation, which is best described on a phonological level, to informational components;

3. with the version of UCG described here, it is not necessary to reduce the arguments of a sign in a fixed order; their order with respect to the head may be underspecified, as tends to be the case, for example, with postmodifier constituents. Such flexibility will allow, for example, a prosodically marked constituent to come after a less marked one, if these appear together in a lexical environment which permits order variation;

4. if the semantic component contains indices, these can be bound to persistent objects in the discourse model, so that it is possible to retain a record of how the latter objects relate to surface expressions used.

Accounts of semantic interpretation are often expressed in terms of a mapping between levels. By using the sign as the basic level of representation, we can maintain a representation of linguistic information simultaneously on a number of levels, from the surface to that of the discourse model. Figure 4.2 illustrates this situation. We may use it to distinguish between the *surface semantics*, or 'propositional form', of an utterance, which is built up according to compositional principles during grammatical derivation, and those discourse entities that parts of it reference. In general,

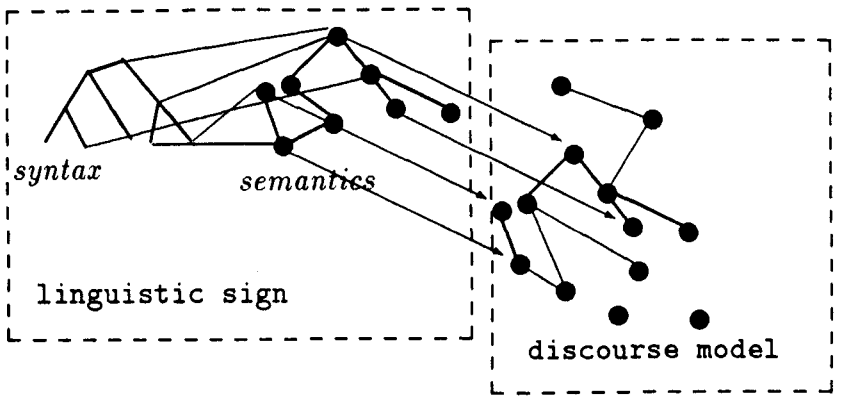


Figure 4.2: Relationship between surface structures, their semantics, and the discourse model

the mapping between the surface semantics and its projection in the discourse model need not be one-to-one.

4.2.3 Representing and updating the linguistic history

The linguistic history may be modelled by a bounded buffer containing entries which combine surface, syntactic, and propositional information, as described above. However, in order for cross-level correspondences to be made for subexpressions within an utterance, it is necessary that not only the sign representing the result of the analysis of an utterance, but also the analysis tree itself, be retained for every entry. The linguistic history is thus updated, in the case of an interlocutor's utterance, by retaining the entire parse tree and adding it as the new entry. In the case of an agent's own utterances, an analysis tree is available as the result of the production process—see Section 5.4. Because a bidirectional linguistic representation is used, the linguistic structure of the agent's utterance may be described in the same manner as that of the interlocutor.

The structure of the analysis tree is described recursively; a non-terminal node corresponds to an instance of functional application, which ultimately derives the string dominated by that node. The immediate daughters of the non-terminal are the head and arguments referenced by that instance of functional application. For details of the analysis tree used in the implementation, see Section 5.4.4. The

linguistic history $\mathcal{L}$ can be described as follows. Given a sequence of utterances, $\langle U_1 \ldots U_k \rangle$, the linguistic history after utterance U_k: $\mathcal{L}_k$, is the sequence of analysis trees $\langle \mathcal{A}_1 \ldots \mathcal{A}_l \rangle$, where l cannot exceed the maximum length of surface recall, $max(\mathcal{L})$, and $\mathcal{L}$ is ordered from most to least recent. Addition to the linguistic history is then as follows: given a linguistic history $\mathcal{L}_k = \langle \mathcal{A}_1 \ldots \mathcal{A}_l \rangle$, and an utterance U_{k+1}, form $\mathcal{L}_{k+1} = \langle \mathcal{A}'_1 \ldots \mathcal{A}'_l \rangle$, by adding the analysis tree corresponding to the new utterance to the front, and dropping the last analysis tree if the resulting length exceeds $max(\mathcal{L})$.

In the discussion that follows, I use the following notational conventions. Given an analysis tree $\mathcal{A}$, $\mathcal{A}.root$ is the sign corresponding to its root; $\mathcal{A}.subs$ are its subtrees, which may be enumerated as the sequence $\langle \mathcal{A}.sub_1 \ldots \mathcal{A}.sub_n \rangle$. The root of any subtree is a sign, so we may refer to its phonology, syntax, and semantics components; for example, $\mathcal{A}.root.phon$ is the phonology of the root sign, $\mathcal{A}.sub_1.root.sem$ is the semantics of the root of the first subtree. Since phonology, syntax and semantics are specified on nodes and not trees, the component *root* may in fact be omitted without risk of ambiguity. I refer to the phonology component of some node within an analysis tree $\mathcal{A}$, as a *subphrase of* $\mathcal{A}$. So for example, after the first three turns of Dialogue 5 from Appendix A, the linguistic history (showing only the *root.phon* components) is as follows:

(4.21)

$$\mathcal{A}_1 = \left[\begin{array}{l} root : \left[\begin{array}{l} phon : London\ to\ Stockholm \ldots \\ \ldots \end{array} \right] \\ subtrees : \ldots \end{array} \right]$$

$$\mathcal{A}_2 = \left[\begin{array}{l} root : \left[\begin{array}{l} phon : I'd\ like\ to\ reserve\ a\ return\ flight \ldots \\ \ldots \end{array} \right] \\ subtrees : \ldots \end{array} \right]$$

$$\mathcal{A}_3 = \left[\begin{array}{l} root : \left[\begin{array}{l} phon : Swedish\ airlines \ldots \\ \ldots \end{array} \right] \\ subtrees : \ldots \end{array} \right]$$

Where turns are multi-sentential, as in this dialogue, a more refined representation

would divide $\mathcal{L}$ on the basis of sentences or phrases, rather than turns. This is in fact done in the implementation described in Chapter 5.

4.2.4 Searching the linguistic history

For the hearer, the computational task of first detecting a case of reuse and then applying the earlier analysis to the present situation may be described as follows. At the time of utterance of the string $U_{k+1} = \mathcal{A}_0.phon$, $\mathcal{L}_k$ consists of a buffer of l previous analysis structures:

$$\langle \mathcal{A}_1 \ldots \mathcal{A}_l \rangle$$

indexed by their strings $\mathcal{A}_1.phon \ldots \mathcal{A}_l.phon$. Starting with $\mathcal{A}_1$, and working backwards, find an analysis tree $\mathcal{A}_i$ which has a subphrase p_i corresponding to some subphrase of $\mathcal{A}_0$. If p_i is coextensive with U_{k+1}, this is a case of *echo*; if p_i matches a subphrase of $\mathcal{A}_0$, I call this *embedded echo.* Alternatively, there may be a number of subphrases of $\mathcal{A}_i$ matching subphrases of $\mathcal{A}_0$; assuming these occur in a similar order in both, then:

1. mark the areas which do not correspond, as *gaps*;
2. form an *abstract* of $\mathcal{A}_i$, by replacing those subtrees corresponding to the gaps with placeholders indicating global and syntactic constraints on them;
3. replace each of these placeholders with the result of analysing the corresponding gap in the newest utterance, $\mathcal{A}_0$.

Since we are dealing with two structures which can be aligned with respect to their similarities, I refer to this case as one of *substitution.* So far as detection of occasions of reuse goes, such an algorithm requires only the phonology components of analysis structures.

For an agent in the role of speaker, the task of generating from a previous entry is twofold: firstly, an appropriate entry which is sufficiently recent must be chosen; secondly, generation should take place in such a way as to maximise use of the

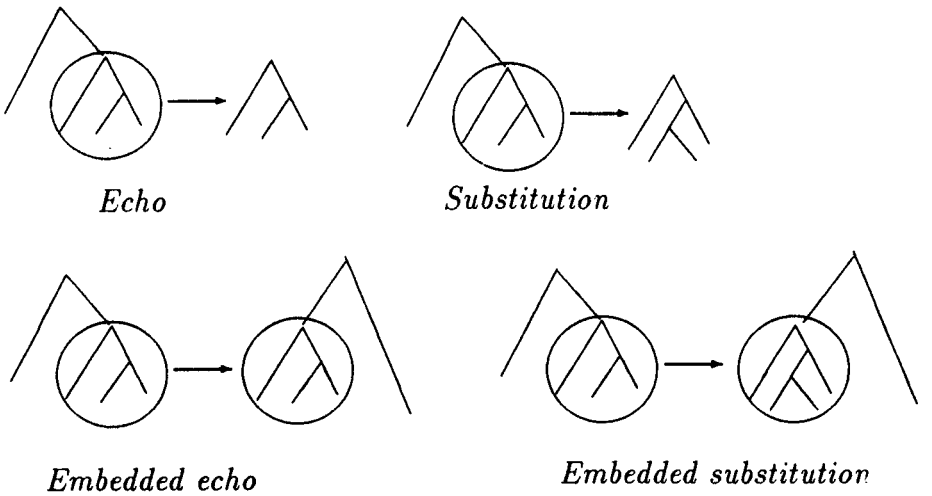


Figure 4.3: Echoes, substitutions, and their embedded versions

previous structure. The ways in which a newly generated sign $\mathcal{A}_0.root$ may draw on a previous analysis tree $\mathcal{A}_i$ are the same as for the agent in the role of hearer, namely:

Echo: The analysis tree $\mathcal{A}_0$ corresponds in its entirety to all or part of the previous structure.

Substitution: $\mathcal{A}_0$ corresponds closely, but not exactly, to all or part of $\mathcal{A}_i$.

Embedded echo: some subtree or subtrees of $\mathcal{A}_0$ correspond exactly to a subtree or subtrees of $\mathcal{A}_i$. structure.

Embedded substitution: This is similar to the case of embedded echo, except that the embedded substructure is not identical to but may be derived via substitution from its antecedent.

Figure 4.3 illustrates the four possibilities. The echoes are a special case of substitutions (where there are no differences); the embedded cases may be handled by beginning the attempt to reuse previous material at some recursive stage within generation. I therefore concentrate on the case where the speaker builds an utterance by applying a substitution to some earlier analysis tree, or a subtree thereof.

Selection of an antecedent structure $\mathcal{A}_i^{sub}$ for a target utterance with semantics $\mathcal{S}$ must be based both on recency and goodness of match. Re-use is of doubtful value

to either speaker or listener, if a more distant rather than a more recent antecedent is used, because of the additional amount of search that this would require. Given a linguistic history $\langle \mathcal{A}_1 \ldots \mathcal{A}_l \rangle$, this means starting with $i = 1$ and searching for an $\mathcal{A}_i$ such that MATCH$(\mathcal{A}_i^{sub}, \mathcal{S})$, where MATCH is a suitably-defined matching relation over pairs of semantic structures.

Having found a candidate for re-use, generation of the target utterance may take place. Two algorithms have been explored. The first one consists in locating those parts of the antecedent analysis tree $\mathcal{A}_i^{sub}$ in advance, and copying over a skeleton analysis tree with these subtrees intact. The skeleton is then traversed in top-down fashion, calling the default procedure GENERATE at those subnodes which are still underspecified.

This algorithm is successful in many cases, where it is possible to lexically instantiate those nodes marked as not recoverable from the antecedent analysis tree. However this is not necessarily the case. For example, assuming that numbers are represented as lists of digits, the algorithm will mark the second parts of [NINE][SEVEN] and [ONE][SEVEN] as the same. This marking will turn out to have been redundant, because the lexical candidate "seventeen" requires no arguments.

An alternative algorithm which combines lexical search with comparison of the source and target structures, assumes that the relation MATCH$(\mathcal{S}_1, \mathcal{S}_2)$ holds for one of three reasons:

1. $\mathcal{S}_1$ and $\mathcal{S}_2$ are identical;

2. $\mathcal{S}_1$ and $\mathcal{S}_2$ have the same semantic type, and furthermore, there is a common lexical entry *lex* such that $lex :< sem >$ is unifiable with either of $\mathcal{S}_1$ or $\mathcal{S}_2$. This means that the two structurally parallel analysis trees will share the same lexical head;

3. similar to (2), only instead of potentially sharing the same lexical head, it is sufficient that there is a $lex_1 :< sem >$ unifiable with $\mathcal{S}_1$, and a $lex_2 :< sem >$ unifiable with $\mathcal{S}_2$, where lex_1 and lex_2 are lexical entries which are semantically close. I shall consider what constitutes semantic closeless below.

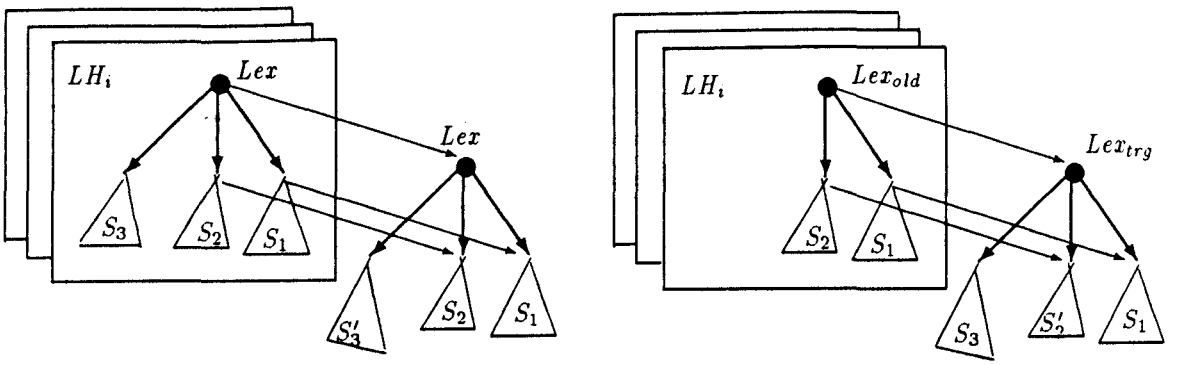


Figure 4.4: Generation of surface forms with lexical head reused or modified

Now given an antecedent analysis (sub)-tree $\mathcal{A}_i^{sub}$ as candidate, the algorithm works by matching $\mathcal{S}_{trg}$, the semantics for the target analysis tree, with $\mathcal{A}_i^{sub} :< sem >$. If the match yields the result *identical*, then the entire antecedent analysis tree is copied.[1] If a common lexical entry *Lex* can be found as the result of the match, then a pairwise correspondence exists between the arguments of *Lex* and those of $\mathcal{A}_i^{sub}$. A recursive call to the algorithm for each of these pairs results in the arguments of *Lex*, $Lex :< args >$ being instantiated either as copies from the arguments of the antecedent, or as modifications of them. The target analysis tree $\mathcal{A}_{trg}$ is then built by combining the information from *Lex* with the information about its instantiated arguments. The situation for matches with reused lexical entry, and for those where a new form must be used is illustrated in Figure 4.4. In the figure the lexical head is shown at the root of its subtree; it is also treated as a terminal, in forming part of the string generated. The first case shown is where the target analysis tree and the antecedent share the same lexical head *Lex*. In that case, the subcategorial arguments correspond, and may have identical values (S_1, S_2) or differ (S_3, S'_3).

If a lexical entry Lex_{trg} for $\mathcal{A}_{trg}$ is not identical but close to that for $\mathcal{A}_i^{sub}$, a correspondence between some arguments may still be possible, and these are recursively generated, as before. However some arguments of Lex_{trg} may not correspond to any from the antecedent; for these it is necessary to apply the default generation

[1] Modulo different morphological instantiation of the head, if this is brought about as a result of different morpho-syntactic features being imposed by the environment. .

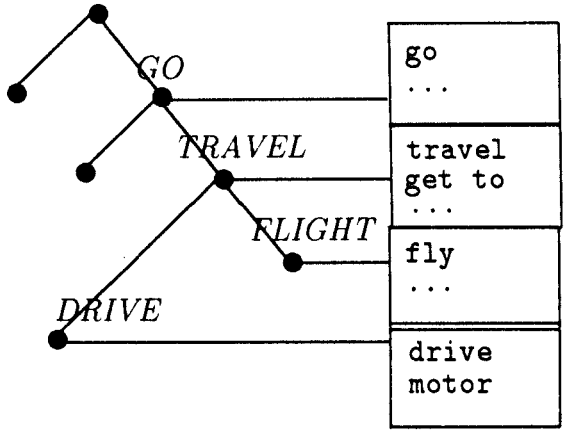


Figure 4.5: Portion of a lexical index showing semantic proximity

algorithm. This second case is also shown in the figure. The new lexical head Lex_{trg} is sufficiently close to the old Lex_{old}. Some arguments still correspond, and have identical values, or differ (S_1, S_2, S_2'); some arguments required by Lex_{trg} however are not present for Lex_{old} (S_3).

To determine matching on the basis of semantic proximity requires that the lexicon be accessible accordingly. Figure 4.5 shows a portion of a lexical index, which includes the lexical items "go", "fly", "travel", and "drive". A lexical entry is attached to the node in the semantic type hierarchy which corresponds exactly to the type of its $<sem>$ component. Ordering among entries at a node is partial. For lexically close entries to match requires additional matching of their arguments. This is illustrated by the case of attempting to generate *seventeen* using the previous structure for *ninety seven.* The roots of the two analysis trees have similar semantics:

$$\left[\begin{array}{l} phon : ninety\ seven \\ syn : head : noun \\ sem : \left[\begin{array}{l} id : Id_{97} \\ type : tens \\ value : 9 \\ next : \left[\begin{array}{l} id : Id_7 \\ type : units \\ value : 7 \end{array}\right] \end{array}\right] \end{array}\right] \quad \left[\begin{array}{l} phon : seventeen \\ syn : head : noun \\ sem : \left[\begin{array}{l} id : Id_{17} \\ type : tens \\ value : 1 \\ next : \left[\begin{array}{l} id : Id_7 \\ type : units \\ value : 7 \end{array}\right] \end{array}\right] \end{array}\right]$$

However the lexical entries for *ninety* and *seventeen*:

$$\left[\begin{array}{l} phon : ninety \\ syn : \left[\begin{array}{l} head : noun \\ args : \langle\ [\ sem : S\ ]\ \rangle \end{array}\right] \\ sem : \left[\begin{array}{l} id : Id_{97} \\ type : tens \\ value : 9 \\ next : S \end{array}\right] \end{array}\right] \left[\begin{array}{l} phon : seventeen \\ syn : \left[\begin{array}{l} head : noun \\ args : [\,] \end{array}\right] \\ sem : \left[\begin{array}{l} id : Id_{17} \\ type : tens \\ value : 1 \\ next : \left[\begin{array}{l} id : Id_7 \\ type : units \\ value : 7 \end{array}\right] \end{array}\right] \end{array}\right]$$

differ in their arguments, and should therefore not be acceptable as candidates for MATCH. In practice however it is simpler to let lexical matching go ahead, and backtrack if subsequently the arguments fail to match.

The algorithm may be illustrated by considering the generation of the phrase: "I drive to Paris", when an analysis tree for "I fly to Paris" is in the linguistic history. First, this entry is chosen, because its semantics (4.22) is similar to that of (4.23):

(4.22)
$$\left[\begin{array}{l} id : flight41 \\ type : flight \\ thetheme : \left[\begin{array}{l} id : speaker \\ type : individual \end{array}\right] \\ thegoal : \left[\begin{array}{l} id : paris \\ type : city \end{array}\right] \end{array}\right]$$

(4.23)
$$\left[\begin{array}{l} id : drive23 \\ type : drive \\ thetheme : \left[\begin{array}{l} id : speaker \\ type : individual \end{array}\right] \\ thegoal : \left[\begin{array}{l} id : paris \\ type : city \end{array}\right] \end{array}\right]$$

However, the lexical head of the former analysis tree, "fly", is incompatible with the target semantics, and there is no other suitable candidate at the node indexed by FLIGHT. So the search continues with candidates nearby in the class hierarchy: first at the node TRAVEL, which is successful but not specific enough, then (successfully) at its daughter node DRIVE. Next, the <syn args> of the instantiated lexical head "drive"

$$\left\langle \left[\begin{array}{l} syn : cat : np \\ sem : \left[\begin{array}{l} id : speaker \\ type : individual \end{array}\right] \end{array}\right] \left[\begin{array}{l} syn : cat : pp \\ sem : \left[\begin{array}{l} id : paris \\ type : city \end{array}\right] \end{array}\right] \right\rangle$$

are matched off against corresponding subtrees in the former analysis tree, on the basis of their semantics. Since an exact match is possible, these subtrees are

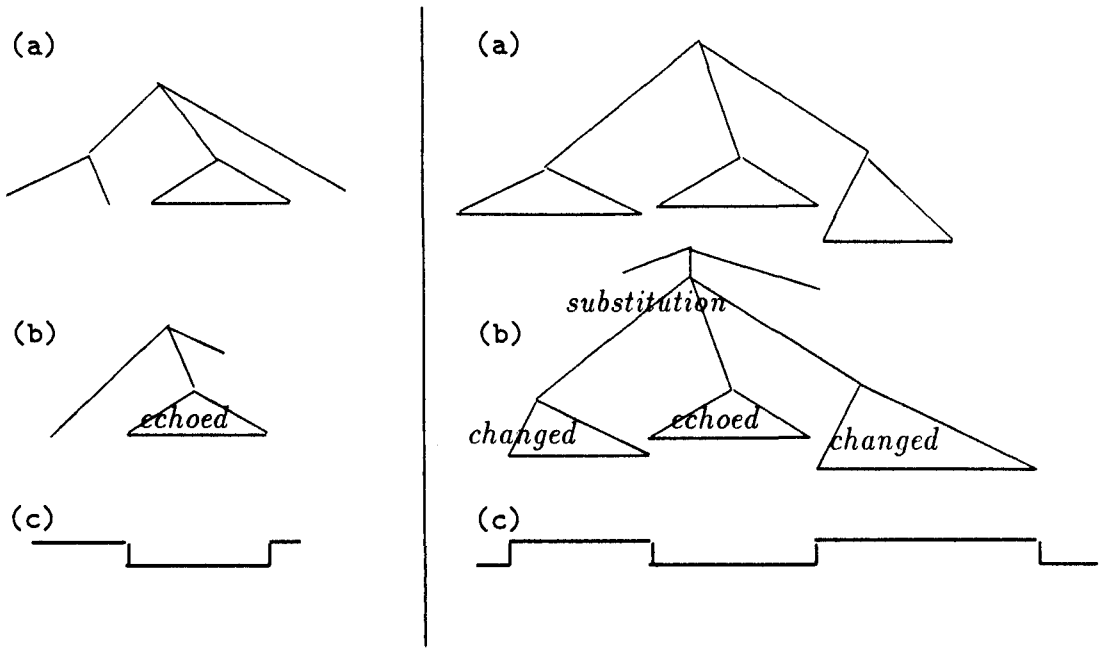


Figure 4.6: Marking scheme for indicating surface reuse

completely copied over onto the target analysis tree.

4.2.5 Prosodic signalling of surface re-use

Echoing and substitution might be indicated on surface structure as follows: constituents which are exact copies of previous structure are marked *s_echoed*; a constituent derived from some previous structure by substitution requires a global marking together with local markings to indicate which parts differ, and which are unchanged. A possible marking scheme is illustrated in Figure 4.6. For the two cases (echo and substitution), structure (a) is the antecedent on which recycling is based; structure (b) is the target. The profiles (c) indicate how a 'naive' prosodic encoding might work. This would depend however on the speaker using stylised intonation, which is not generally the case when information transfer is intended. If instead we take sentence accents to be local, marked on syllables, then the listener is at least able, after lexical segmentation, to interpret these accents as markings on words. Elements marked *changed* could be distinguished from constituents which are simply new, by greater pitch prominence. Echoed constituents however are not inevitably 'de-accented' in the sense that they lack any accent; as Horne (1991) demonstrates,

pre-nuclear constituents in particular tend to have weak accents, even when they echo given material. Only when they occur within the nuclear tail are constituents necessarily without accent. Some form of relational representation is therefore to be preferred to one which is strictly binary, or three-valued.

Signals based on patterns of prosodic prominence only serve to indicate how the current structure compares with previous ones. Such prominence information is no longer useful when the utterance in question is no longer the current one. For this reason, I assume that prominence information is not retained when an utterance is added to the linguistic history.

4.2.6 Reuse and ellipsis

The proximity of former analysis structures required for the markings *s_modified* and *s_echoed* to be generated, raises the possibility of the speaker using ellipsis as an alternative. Consider for example, the following:

(4.24) *A*: you want to fly to manchester
B: no. I want to < *drive* > to manchester

(4.25) *A*: you want to fly to manchester
B: no. drive.

According to the Gricean maxim of quantity, (4.25) should be preferred, all other things being equal. Both speaker's production algorithm and listener's search algorithm can be modified, as follows: speaker—generate as for a substitution utterance, but only utter the part(s) marked *s_modified*; listener—assume this is a substitution utterance (in Example 4.25 the assumption is warranted by the presence of the discourse marker "no"), locate a former (sub-) tree in the linguistic history, whose semantics is in the proximity of the head semantics of the utterance, and perform the necessary substitution. [2] In the case of echo utterances, the search is even more

[2]The apparent anomaly in (4.25), whereby instead of ellipsed *you* in *B*'s utterance, we are supposed to read *I*, can be resolved by the requirement that common portion match semantically as well as syntactically, with semantic matching taking precedence. Semantic matching is discussed in the following section.

straightforward; a substring (or substrings) from the previous analysis structure should exactly match the current utterance:

[37] T2:SA:1235 (T)

(4.26) *A*: yes that's arrived at eleven oh five
C: eleven oh five

In (4.26), *A* is able to infer that *C*'s utterance matches her own, and therefore may be expanded to have the full semantic representation as that. A pragmatic interpretation of its dialogue function can then be made. Routinely, this will be a confirmation request. In some cases, the intonation may serve as a cue that a non-routine interpretation is required:

[2a] T1:SA: (T)

(4.27) *C*: well he's got four forty five leaving malaga
A: (well it's not our ...)
H leaving $^{\uparrow\uparrow}$**HL** malaga

Elliptical references to the previous utterance are extremely common, especially when confirmation is being sought or given. However, there are many cases from the Flight Enquiries corpus where ellipsis could have, but has not, been used, for example:

[5] T1:SA:1356 (T)

(4.28) *C*: which terminal will I come back to
A: you'll come back to north terminal

[7] T1:SA:2013 (T)

(4.29) *A*: you say it's arriving on the [@:m]
C: it's arriving early morning

[3] T1:SA:632 (T)

(4.30) *A*: yeah it's definitely in the zone though
CC: it's definitely in the zone at the moment

This indicates that other factors than Gricean economy may be at stake. In each example, information is being added: *north terminal, early morning, at the moment*; providing a frame of reference which the listener can be assumed to still have access to may facilitate semantic processing, which will require the location of the new information within the frame of reference.

4.3 Focus in the discourse model

Evidence that speakers maintain internal discourse models is necessarily indirect. However the assumption that speakers add to and reference internal models greatly facilitates the analysis of a discourse. Consider the following example:

(4.31)

[38] T2:SA:1516 (T)

3	C:	good afternoon I'd just like
4		to confirm a flight tonight
5		ahm (.) supposed to be flying to cyprus
6		on bee ay six six eight at ten o'clock
7		(1)
8	A:	right I'll check that for you
9	C:	thank you
10		(4.3)
11	A:	to larnaca
12	C:	that's right yes
13	A:	yes: that's: er scheduled for take off
14		at twenty two hundred
15	C:	right and uh what time am I
16		supposed to be at the airport
17		to check in
18	A:	yes about two hours before departure
19	C:	two hours
20	A:	yes
21	C:	a:nd (.) which terminal is that
22	A:	it's the north terminal at gatwick

C's initial utterance introduces a discourse entity corresponding to a flight, which is related to the speaker (as passenger), and to a place time, and flight number. *A* is able to infer a default destination airport of Larnaca, although this has not been mentioned, because she can assume that she and her interlocutor have a common internal representation of the island of Cyprus. Similarly, references to the take-off event (by *A*) and the check-in (by *C*) indicate that they can be assumed to be in the respective interlocutor's model. The dialogue also illustrates how anaphora are used to refer to discourse entities which are temporarily in focus: "that" in lines 8 and 13, refers to the flight; in line 21, it is used to refer to the check-in event.

Speakers construct mental models of the situation in a cooperative and sympathetic way. They need to be aware of what default knowledge on the part of the interlocutor it is safe to assume—that Larnaca is the main airport in Cyprus, for example. When these assumptions break down, repair is always possible, as in the

continuation of the above dialogue:

(4.32)

[38] T2:SA:1516 (T)
24 C: ...does the railway
25 take me to the north terminal
26 A: I beg your pardon
27 C: the railway take me to
28 the north terminal
29 (.3)
30 y'know the trains at gatwick station
31 A: (u)hmm ituh
32 I think it just takes you into gatwick

A indicates incomprehension, and *C* makes a double attempt at repair: firstly, to repeat himself, then to expand the context in which "railway" is to be interpreted, incidentally introducing more discourse referents corresponding to "trains" and "Gatwick station".

It is useful therefore to distinguish a number of levels at which an agent may maintain an internal model of the discourse. Firstly, the discourse model may contain only those discourse entities which are referred to explicitly by one of the participants, or which are deictically present. This is the definition of 'discourse model' that a number of authors use, for example Levelt (1989), for whom the discourse model consists of the union of the speaker's and interlocutor's contributions. Secondly, the discourse model may contain that extension of the world of those elements which have been explicitly introduced, which is necessary for coherence. I have illustrated the working of this with respect to the Gatwick–Cyprus example. This demonstrates also that an agent does not necessarily extend the model in all its full details—even if he had the mental capacity to represent an entire world internally, this would run the risk of departing from the interlocutor on too many details. Instead the situation which the agent represents to himself is a partial model of the actual situation, with many details left unspecified. The agent, for example, does not initially include the train station in her representation of Gatwick airport, and in fact has some difficulty at first in doing so. Finally, an agent may represent to herself facts and assumptions about the interpersonal situation which only ever get referred to indirectly, if at all.

I shall adopt the second definition of a discourse model, namely, the plausible extension of the situation represented in the discourse so far, which an agent adopts and refers to in his utterances.

This section is organised as follows: in (4.3.1) I outline a knowledge representation formalism for representing structures within the discourse model, and semantic information generally. In (4.3.2) a number of mechanisms are proposed according to which discourse entities in particular and semantic entities in general may be ordered for prominence. Section 4.3.4 discusses the particular case of contrastive focus, especially that which comes about when repair utterances are employed to convey and acknowledge modified information states. To handle these cases a new representational device, of accessing layered information states within the discourse model via world-indices, is introduced; this is then applied to the case of alternative solutions.

4.3.1 Representing information in the discourse model

4.3.1.1 Knowledge representation

Representing discourse entities and their interrelationships can be done in a relatively straightforward manner using a *semantic network* (eg Brachman and Schmolze 1989). Unlike representations based on predicate calculus, networks have the advantage of encoding straightforwardly the persistence of objects. They are also amenable to graphical representation. Monotonic addition of information can be effected by simply growing the relevant portion of the network to include the new entities and relations (see Figures 4.7 and 4.8 for examples).

The choice of semantic primitives for the knowledge representation language is to some extent arbitrary. The representation used here is based on typed discourse entities, related to one another via semantic *roles*. The discourse model at any one time consists of whatever such instances currently exist, together with the relations between them. Discourse entities (or objects) correspond not only to the kind of thing referred to by nominal expressions in English. I follow usual semantic network practice, and recent literature on linguistic semantics (eg Calder et al. 1988a, Hobbs

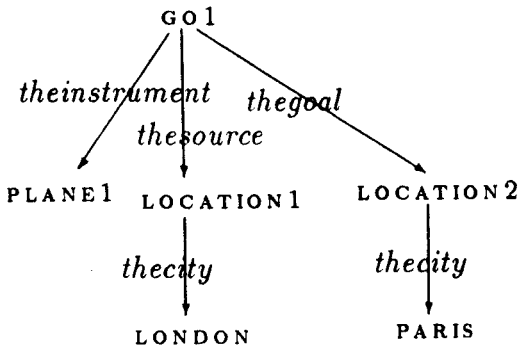


Figure 4.7: Discourse model after utterance: "fly from London to Paris"

1985) and *reify* verbs—ie, represent the relevant events, states, and relations as tokens denoting discourse entities. This greatly facilitates representation, and may be partially argued for on grounds of anaphoric usage; consider, for example the following:

(4.33) *A* i want to travel to Perros
B sorry. where is (the journey/that) to?

To the extent that nouns and verbs reference the same event, they may be represented as the same discourse entity. In (4.33), "travel" and "journey" might both refer to the same discourse entity, SINGLE_JOURNEY1.

Using a semantic network representation, I illustrate how the discourse model is extended over two successive inputs:

(4.34) fly from London to Paris
(4.35) travelling from Heathrow at 17:15

As a result of utterance (4.34), with semantic representation:

(4.36)

$$\begin{bmatrix} type : go \\ theinstrument : [\ type : plane\ ] \\ thesource : [\ thecity : london\ ] \\ thegoal : [\ thecity : paris\ ] \end{bmatrix}$$

the discourse model contains the information shown in Figure 4.7. The semantic representation of the utterance in (4.35) is:

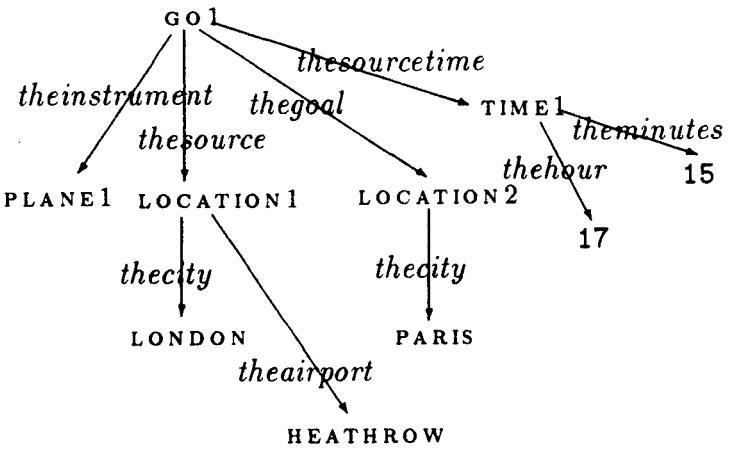


Figure 4.8: Discourse model after the utterance: "travelling from Heathrow at 17:15"

(4.37)

$$\left[\begin{array}{l} \textit{type} : \textit{go} \\ \textit{thesource} : \left[\ \textit{theairport} : \textit{heathrow}\ \right] \\ \textit{thesourcetime} : \left[\begin{array}{l} \textit{thehour} : 17 \\ \textit{theminutes} : 30 \end{array}\right] \end{array}\right]$$

As a result, the discourse model gets extended as shown in Figure 4.8. This extension assumes that we can infer that "fly" and "travelling" refer to the same event.

The power of representation may be further extended by the addition of *rules of inference*, or constraints. For example, world knowledge tells us that flying entails a journey, and that the latter has arrival and departure events associated with it. It would then be possible to extend the utterances in (4.34–4.35) by the inputs:

(4.38) leaving from terminal 4

(4.39) arriving at Charles de Gaulle airport at 1730

Figure 4.9 shows the results of operating an inference rule that adds the discourse events corresponding to JOURNEY, ARRIVAL and DEPARTURE. If all applicable constraints have been imposed, I call the resulting state of the discourse model *inferentially complete.* Explicitly mentioned entities may need to be distinguished from inferred ones, in case referring to the latter is done in a different way—for example they may be presented as being less accessible.

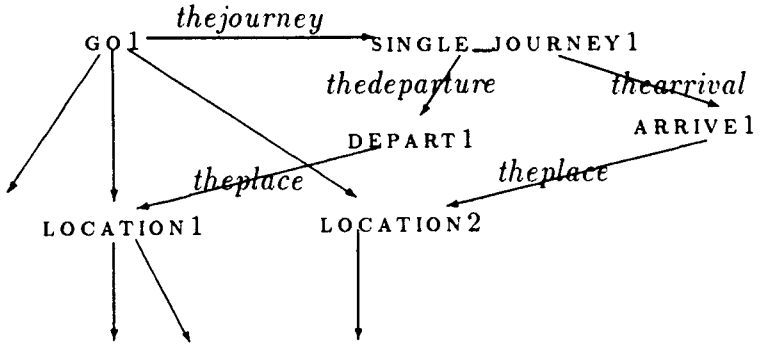


Figure 4.9: Discourse model extended by inferred entities and roles

4.3.1.2 Representing semantic input and output

It is useful to distinguish on the one hand the informational content of an utterance viewed as a set of constraints encoded in a subgraph of the discourse model; and on the other the *semantic representation* which is compositionally derived from a natural language expression, with the aid of the lexicon. Assuming that interpretation is modular, this representation is not initially anchored in context; the semantic representation is ambiguous as to what discourse entities are being referred to. Expressions of the semantic representation language, however, make use of the same primitives as those used in describing the contents of the discourse model. However the representation is closer to predicate logic, allowing in particular the possibility of quantifiers, determiners, disjunction and negation. Using the binary-valued feature DEF to denote definiteness, for example, we can represent: "the flight from paris is delayed" as follows:

$$\left[\begin{array}{ll} \textit{type} : \textit{delay} & \\ \textit{thetheme} : & \left[\begin{array}{l} \textit{type} : \textit{single\_journey} \\ \textit{def} : + \\ \textit{thedeparture} : [\ \textit{theplace} : [\ \textit{thecity} : \textit{paris}\ ]\ ] \end{array}\right] \end{array}\right]$$

In interpreting this input as an assertion, it would be necessary to find out to what extent it asserted new information, and to what extent it simply referenced existing information. In this example, SINGLE_JOURNEY and PARIS are existing information, and need to be linked to existing discourse entities. DELAY is new information, requiring the addition of a new discourse entity, with the appropriate links. In many European languages, where a common noun is used to describe a discourse entity, whether or not it is assumed to be already present in the discourse model is signalled by the use of a definite or indefinite determiner, respectively.

When interpreting input, definiteness markers and other accessibility markers such as pronouns are used to guide the search for an anchoring to some existing discourse entity. Once found, however, they are not retained as part of the discourse model's information.

Input is therefore interpreted as being about the discourse model. If an interlocutor's utterance is pragmatically interpreted as an assertion, information is added. In the cases of queries about the content of the discourse model information may not be added, although it can be, as in the following example:

(4.40)
C i want to fly from London to Paris
A what time do you want to leave Heathrow

—where *A* has made the default inference that the departure airport is heathrow.

Because the same linguistic knowledge base is being used for generation as for parsing of input, descriptions to be generated by the Agent also have to be planned in terms of the semantic representation language. To do this requires both ordering (since representations in the discourse model are on the face of it unordered), and the ability to draw on context and make anaphoric references where possible.

I shall commonly use the name of a semantic type in small capitals to refer to a semantic structure of that type. Thus for example

$$\begin{bmatrix} id : go1 \\ type : go \\ thesource : [\ \ldots\ ] \\ \ldots \end{bmatrix}$$

may be referred to simply as GO, or GO1 if it makes a difference which semantic structure of type *go* is being referred to. Where it is clear that I am talking about discourse entities viewed as nodes in the semantic network rather than about entire structures, I shall use the same shorthand to denote the nodes. An extension of this shorthand is to use strings indicating values to stand for structures containing information which may be complex. Thus for example, 930 stands for:

$$\begin{bmatrix} id : \iota \\ type : time \\ thehour : \begin{bmatrix} type : hour \\ value : 9 \end{bmatrix} \\ theminutes : \begin{bmatrix} type : minutes \\ value : 30 \end{bmatrix} \end{bmatrix}$$

In a similar fashion, I use expressions like SA790 to refer to flight numbers, and LUTON to refer to cities.

4.3.2 Accessibility and prosodic focus

The notion of accessibility (cf. Section 2.3.1) is one that has emerged largely out of consideration of discourse coherence mechanisms. Using the plausible metaphor of levels of activation over a memory network, the accessibility principle treats more active discourse entities as more accessible to an interlocutor. Relative accessibility is then encoded linguistically using a number of devices, including: word order, pronominalisation, prosodic (de-)focussing. This thesis is largely concerned with the last of these; however a theory of the effects of accessibility on prosodic focus ought to be compatible with accounts of its influence on the pronominalisation decision, and should explain clashes where they occur (see Section 4.3.3.3 for examples).

Aspects of the speaker's message which appear to exhibit given-new information structure are often predominantly studied with reference to the temporal organisation of discourse. However, as Bolinger (1989) points out, a speaker may also take into account his own momentary assessment of the relative *interest* value of discourse entities.

In an attempt to review the evidence concerning accessibility and prosodic focus, I first present an analysis of the Swedish Airlines corpus. This material was designed for the purpose of investigating prosodic behaviour from an interactional point of view, rather than specifically to look at discourse referents. Nevertheless the dialogues exhibit a number of the qualities of authentic spontaneous material which render a study of accessibility possible, including a considerable amount of redundancy and repetition.

For each dialogue, a number of nominal and verbal groups were singled out, on the basis that they were deemed to refer to existing material—either mentioned or inferrable. An example dialogue is given in (4.41):

	Word	Accent Score/4	Distance	Verbal	Intentional
SA8:C1:	flight	0	–	–	–
SA8:A2:	london	4	1	–	TP
SA8:A2:	stockholm	4	1	–	TP
SA8:A3:	flight	1	3	–	RF
SA8:A3:	london	4	3	–	RF
SA8:A3:	stockholm	4	3	–	RF
SA8:A3:	sunday	1	3	–	RF
SA8:A3:	morning	4	3	–	RF
SA8:C3:	flight	0	2	–	–
SA8:C3:	morning	3	2	–	–
SA8:A7:	flight	2	7	–	–
SA8:C7:	that	1.5	1	–	–
SA8:C7:	morning	0	8	–	–
SA8:A9:	reserve	1.5	–	+	–
SA8:A9:	seat	4	–	–	–

Table 4.1: Results for SA dialogue 8, showing deaccenting scores

(4.41)

SA8:C1: is there a flight on sunday morning, london to stockholm
SA8:A2: london to stockholm
SA8:C2: sorry I didn't get that
SA8:A3: you want a flight from london to stockholm on sunday morning
SA8:C3: that's right
SA8:C3: is there a flight around mid morning
...
SA8:A7: well,there's a flight at eight fifteen
SA8:C7: is that the only one in the morning
SA8:A8: yes
SA8:C8: okay then
SA8:A9: do you want to reserve a seat
...

For each of the four speakers, the selected phrases were given scores of 0 or 1 according to whether or not they were defocussed. In case of an uncertain judgement, a fractional score (typically 0.5) was given. The results for the four speakers were pooled, so that each target phrase received a score out of four, and tabulated as in Table 4.1. The distance from the antecedent is a measure (in turns) of the distance between the token, and some earlier expression assumed to be co-referential with it. Where no distance is given, the entity is assumed to be potentially inferrable from context. For example, *reserve* or *seat* might be taken to be inferrable, if the service is known to be concerned with flight bookings. The label *Intentional* is discussed below. Results discussed below use the combined data of all nine dialogues; tables

	Near	Mid	Far	Inf
No Filter	2.71	2.58	3.11	1.81
Filter	1.07	2.29	2.39	1.25

Table 4.2: Distance and focussing score: all dialogues

for these are given in Appendix C.

The first question concerns the relationship between distance and de-focussing. Marked tokens were divided into four groups, *NEAR*, *MID*, *FAR* and *INF*, as follows: *NEAR* was the group with *Distance* ≤ 1, *MID* had $2 \leq$ *Distance* ≤ 5, *FAR* had *Distance* ≥ 6, and *INF* was the group of inferrables. Table 4.2 shows the average scores for these groups. In the condition *No Filter*, all results were considered together. Assuming a score of 2 below which a word can be considered defocussed, all mentioned tokens are relatively focussed; only inferrables may be said to be defocussed. However, as I propose in Section 4.4, some discourse entities, notably those intentionally mentioned as task parameters, need to be absolved from rules about accessibility and deaccenting. I therefore marked such words *TP*. In addition, task responses ($\langle$*resp*; *task*$\rangle$) and reformulations ($\langle$...*repeated* : +...$\rangle$) were marked *RQ* and *RF* respectively; utterances of both kinds could be considered to involve accents of power. The result, shown as *Filter* in Table 4.2 indicates a progression of increasing focus from Near to Mid to Far, with Near defocussed. These results suggest a falling off of accessibility with time; however, this need not be so. In Table 4.1 for example, [SA 8:C3]:*morning* was deaccented by only one speaker, whereas [SA 8:C7]:*morning*, at a comparably greater distance from its most recent antecedent, was deaccented by all speakers.

The tables were examined for evidence of defocussing of verbal expressions. The results are shown in Table 4.3. Verbs with score less than 2 were classed as defocussed; otherwise they were focussed. When *RF* and *RQ* are not filtered out, there is no clear preference for defocussing verbs. When they were, there was a tendancy for verbs not previously mentioned to be defocussed. An explanation for this could be that verbs which are classed as inferrable are relatively uninteresting, or less informative compared to other material in the same phrase.

Consideration of those tokens treated by all speakers as defocussed reveals a

All utterances		
	Focussed	Defocussed
Mentioned	6	3
Not mentioned	8	7
Filtered		
	Focussed	Defocussed
Mentioned	4	1
Not mentioned	3	7

Table 4.3: Defocussing among verbs: all dialogues

number of scenario-dependent tokens: *flight, travel, arrive.* These when they occur are relatively uninteresting; in Bolinger's words (1986), accent gets "sacrificed to nearby focal meaning". How is it then that the token *three* in [SA 9:A4'], which is undeniably part of a task parameter, and new information, also gets deaccented? The most likely reason is the occurrence of *three* in a structurally similar position in the previous phrase. Structural parallelism has already been discussed in Section 4.2; many of the arguments concerning the processing advantages of exploiting it can equally be applied at the discourse level.

This study of the Swedish Airlines corpus has been of some exploratory value, suggesting that there is a relation between defocussing and recency, and pointing out the importance of inferrability, and its dual relative informativeness. In many cases, the default focus pattern due to accessibility may be overruled by the speaker's intention. In Section 4.3.3 I present a computational account of prosodic focus in discourse which extends these results. In it I attempt to disentangle such factors as recency, mention, deixis, inferrability and relative informativeness. The theory of relative focus that I present derives in part from Bolinger's notion of *interest* (Bolinger 1986, Bolinger 1989). Excluding for the present intentional focussing (Section 4.4), the following factors are proposed:

1. mention, either exactly or in modified form, in previous discourse;

2. inferrability—specifically that of scenario-dependent entities;

3. deixis—in particular the accessibility of certain privileged entities that form part of the current discourse situation;

4. relative informativeness within the current structure.

Unlike factors 1) and 2), 3) and 4) have not been explicitly investigated in this section. Nevertheless, previous studies point to their importance (see Sections 2.5.1–2.5.2).

4.3.3 Modelling accessibility for prosodic focus

4.3.3.1 A relational representation of prosodic focus

Labelling word tokens according to a binary valued feature *focus* may be overly restrictive. Take the utterance:

(4.42) SA7:A4: travelling from Stansted

Both *travelling* and *Stansted* are accented by all speakers; however, the labelling scheme fails to account for the considerably greater prominence of *Stansted.* Pre-nuclear sentence accents often count for little, their presence being attributable to the relative stability of the 'hat pattern' (Bolinger 1986), rather than to any necessary focus on the accented material. The power of accents generally increases from left to right, culminating in the nucleus, with any material following it usually deaccented. To account for this I propose a relational representation that divides the discourse entities in an utterance into more or less prominent.

I define a partial ordering of prominence over discourse entities. The ordering is partial because the relative prominence of an arbitrary pair of discourse entities may be indeterminate. I represent the ordering relation with the operator '$\prec$'. So for the utterance

(4.43) >you're< <travelling> from <Paris>

the ordering on discourse entities can be represented:

(4.44) CALLER $\prec$ {GO, PARIS}

Note that the delimiters < ... > and > ... < marking textual examples are not capable of making all the distinctions implicit in the ordering of semantic entities. To do this properly requires that an ordering be established over non-terminal constituents—see below.

When it comes to representing abrupt discontinuities in prominence, such as are present in emphatic stretches of text, the ordering relation '$\prec$' is insufficient. I therefore define also the ordering relation '$\prec\prec$'. A variant of Example 4.43 with emphatic focus on *Paris* is the following:

(4.45) >you're< <travelling> from ≪Paris≫

with the corresponding ordering over semantic indices:

(4.46) CALLER $\prec$ GO $\prec\prec$ PARIS

Relations (4.44) and (4.46) are compatible, the latter refining the former.

A proper derivation of metrical prominence takes as its input a partial ordering over semantic entities, and produces a metrically-labelled binary tree, of the kind discussed in Giegerich (1985): that is, all nonterminal nodes are binary-branching, with their sub-branches labelled strong (s) and weak (w). Because lexical entries contain a semantic component, once the surface structure has been generated, every semantic element in the ordering can be associated with a surface constituent. This is normally situated where the lexical entry was put into the tree. A prominence order over surface constituents may be converted into a prosodic representation, for example a metrical tree. An algorithm for doing this would iterate over the semantic entities specified for prominence, in descending order, at each stage if necessary manipulating the relative strength marked of subtrees so that the designated terminal element (DTE) of those portions of the tree as yet unassigned is associated with the currently most prominent semantic entity. In the case of underspecification of semantic prominence, or no specification, metrical ordering defaults to that required lexically (eg by entries for 'frozen' noun compounds), or to the even more basic default of prominence to the right.

4.3.3.2 Historically derived focus

I consider the possible relations between the current semantic form and members of an *accessibility history* of past semantic forms. This idea is similar to that investigated in Section 4.2.3, but whereas there complete records of linguistic structure

were being compared, and primarily from a lexical point of view, in this section only the semantic components of utterances are under investigation. Again, the length of history available at any given time is trimmed to a fixed upper limit, so that a weak notion of recency is operative. When they reach a certain age (determined by this limit), items are discarded.

At first, it might seem appropriate to keep only a record of discourse entities. The accessibility history however contains entire structures corresponding to past utterances. The pattern of prominences derived may thus reflect modifications between structures; subsequently the listener will be able to use this information to build a representation of the current structure on the basis of the previous one. Turns $A1$ and $A3$ of the dialogue in (4.47) illustrate how this can happen:

(4.47)

C1	I want to go from London to Paris
A1	what time do you want to arrive in >Paris<
C2	two thirty
A2	two thirty
C3	that's right
A3	(and) >what time< do you want to <leave> <London>
C4	>leave< >London< <about noon>
A4	sa 123 >leaves< >London< at twelve thirty
A4'	and <arrives> in <Paris> at <two thirty>

An acceptable pattern of prominences for $A1$ is:

(4.48) <what time> do you want to <arrive> in >Paris<

ie, the arrival event and the unspecified time are given increased prominence. In uttering $A3$, however, the Agent may organise the prominences to the effect that the parallel with $A1$ is brought out:

(4.49) >what time< do you want to <leave> <London>

Here it would appear that the two utterances are matched at the outermost (structural) level. Retaining entire structures also means that when a (sub)expression $\mathcal{S}$ corresponds exactly to a previous structure or to some subexpression of it, it is sufficient to mark this sharing at the root node of $\mathcal{S}$, rather than marking all its nodes.

Let us now consider the dialogue (4.47) in more detail. A reduced form of the semantic structures kept in the accessibility buffer is shown in Figure 4.10. The

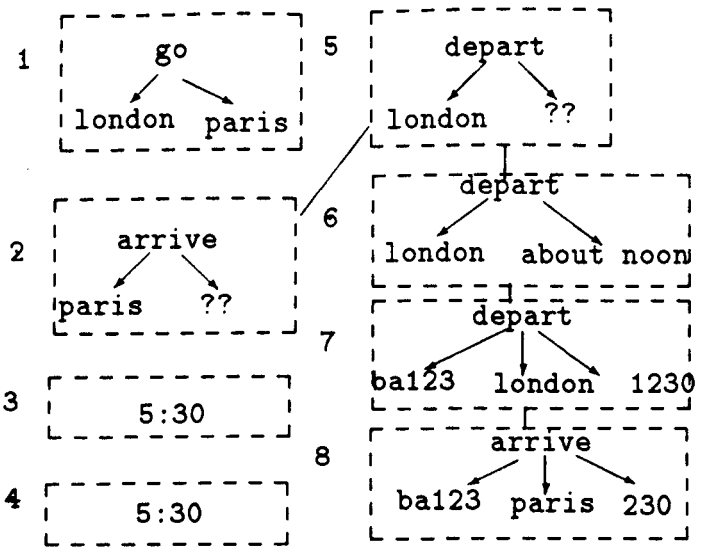


Figure 4.10: Buffer for accessibility history dialogue

evolution of the accessibility buffer over time is straightforward. At the time of $C1$ it is empty; at the time of $A1$ the contents are [1]; at the time of utterance $C2$ the contents are [1,2]; and so on, until by the time of $A4'$ the contents are $[8-n,\ldots,7]$ for a maximum buffer size of n ($n \leq 7$). No representation is shown for $C3$, which lacks a referential component. Nor is WANT represented, because in the simple account of semantic interpretation presented here, propositional attitude operators are not represented within the discourse model. I return to their treatment in Section 4.3.3.5.

An algorithm concerned with deriving patterns of focus via the accessibility history must account for two phases: matching the current structure or some of its substructures with previous structures or substructures, and as a result of these matches, deriving the ordering relation over the current structure. I first assume a matching algorithm, and consider how the results can be arrived at. I use the notation $[n \leftarrow m]$ ($n > m$) to refer to the situation where structure n in Figure 4.10 or a substructure thereof is matched against structure m or a substructure of it.

$[2 \leftarrow 1]$ no acceptable modification at top level will derive 2 from 1. PARIS is common, so it has reduced prominence.

$[3,4 \leftarrow \ldots]$ nothing interesting happens here; items $C2$–$C3$ only appear in the dialogue to demonstrate that $A1$ and $A3$ are not required to be adjacent.

[5 ← 2] DEPART and ARRIVE are closely semantically related—they have a common supertype (cf. Figure 5.4). They also have matching roles: *theme*, *thetime*, *theplace*. TIME is common[3], the other elements correspond to modifications, giving TIME ≺ LONDON, DEPART.

[6 ← 5] The structures match at root level, and have *theplace*:LONDON in common: LEAVE, LONDON ≺ ABOUT NOON.

[7 ← 6] Similarly: LEAVE, LONDON ≺ SA123 ≺≺ 1230. 1230 differs from ABOUT NOON, so has heightened prominence.

[8 ← 7] Finally, we obtain for [8]: SA123 ≺ ARRIVE, PARIS, 230. In the string uttered, the second occurrence of SA123 is ellipsed by the rule for vp-conjunction, which is implicit in the lexical entry for *and*.

An algorithm capable of accounting for these patterns of prominences can be summarised as follows:

1. Given a current structure with semantic representation $\mathcal{S}_0$, and an accessibility buffer $\langle \mathcal{S}_1 \ldots \mathcal{S}_n \rangle$, where $n \leq max$, the maximum buffer size, a structure to match $\mathcal{S}_0$ is found by searching backwards for a structure $\mathcal{S}_i$, or some substructure $\mathcal{S}_i^{sub}$ which matches $\mathcal{S}_0$ entirely. If none is found, start again with some substructure of $\mathcal{S}_0$, $\mathcal{S}_0^{sub}$, searching backwards for some structure which matches that in its entirety. So for example, with [2 ← 1], $\mathcal{S}_0$ is [2], $\mathcal{S}_i$ is [1], and matching doesn't succeed until the level PARIS is reached—ie, we have a match between $\mathcal{S}_0^{sub}$ and $\mathcal{S}_1^{sub}$. With [5 ← 2] on the other hand, with [5] corresponding to $\mathcal{S}_0$, and [2] to $\mathcal{S}_i$, matching happens at top level. The order of search is important; searching isn't allowed on behalf of substructures of $\mathcal{S}_0$, until it has been tried at root level, for the whole buffer. Otherwise, there would be nothing to prevent the structure TIME matching with '230' from [4].

2. Having located a matching pair $\langle \mathcal{S}_0^{sub}, \mathcal{S}_i^{sub} \rangle$, the resulting prominence pattern is determined thus:

[3] In fact identifiers representing unknown discourse entities of the same type need to be kept distinct, as they will probably be different once instantiated. The algorithm therefore allows unspecified discourse entities of the same type to match.

If $\mathcal{S}_0^{sub} \sqsupseteq \mathcal{S}_i^{sub}$, ie, $\mathcal{S}_0^{sub}$ is an extension of $\mathcal{S}_i^{sub}$, as for example: [*type* :DEPART, *thedest*: PARIS] is an extension of [*type* :DEPART], then $\mathcal{S}_i^{sub}$ marks the echo part, giving

(4.50) $$\mathcal{S}_i^{sub}.ids \overset{*}{\prec} \mathcal{S}_0.ids - \mathcal{S}_i^{sub}.ids$$

where '–' denotes set difference, and '$\overset{*}{\prec}$' is the relational operator '$\prec$' extended for sets, ie

$$A \overset{*}{\prec} B \triangleq \forall\, a_i \in A, b_j \in B \quad (a_i \prec b_j)$$

If $\mathcal{S}_0^{sub} \sqsubset \mathcal{S}_i^{sub}$ then

(4.51) $$\mathcal{S}_0^{sub}.ids \overset{*}{\prec} \mathcal{S}_0.ids - \mathcal{S}_0^{sub}.ids$$

I use $\mathcal{S}.ids$ to represent the discourse entities contained in the structure $\mathcal{S}$. Henceforth I drop the *ids* component, where context makes it clear that we are referring to a set of discourse entities, and not a structure.

Otherwise, defining *Common*, *Changed*, *New* as in (4.52–4.54), the case where neither of the parallel substructures subsumes the other is given in (4.55).

(4.52) $$Common \equiv \mathcal{S}_0^{sub} \cap \mathcal{S}_i^{sub}$$

(4.53) $$Changed \equiv \mathcal{S}_0^{sub} - Common$$

(4.54) $$New \equiv \mathcal{S}_0 - \mathcal{S}_0^{sub} - \mathcal{S}_i^{sub}$$

(4.55) $$Common \overset{*}{\prec} New \overset{*}{\prec} Changed$$

That is, corresponding elements which are different are most prominent; common elements are least prominent, with the remainder in between. This is illustrated in [5 ← 2], with *Common* = TIME and $\overline{\mathcal{S}_0^{sub}}$ = {ARRIVE,PARIS}. These two sets together partition the semantics of the utterance, so that the middle term is absent. The effect of the rule is to raise in prominence changed elements and lower elements which are shared by the two structures.

I use the notation $\prec_h$ to represent ordering of discourse entities derived with

respect to the accessibility history. I drop the distinction between $\prec$ and $\overset{*}{\prec}$; which of these is intended can usually be made out from the context. In expressions of the type $A \prec REST$ or $REST \prec A$, $REST$ should be taken to denote the complement of A with respect to the total set of entities under consideration. Thus (4.50, 4.55) may be rewritten:

(4.56) $$S_i^{sub} \prec_h REST$$

(4.57) $$Common \prec_h REST \prec_h \overline{S_0^{sub}}$$

An example of rule (4.56) is found in the Cyprus dialogue discussed in Section 4.3:

(4.58)

[38] T2:SA:1516 (T)
22 A: it's the north terminal at gatwick
23 (.)
24 C: ...does the railway
25 take me to the north terminal

Not only is the unfocussed reading > *north terminal* < preferred; to focus this phrase would create the unwanted presupposition that C had not heard the earlier mention, or wished it to be assigned special prominence. Rule (4.57) is illustrated in (4.59):

(4.59) my uncle inherited a fortune,
but my aunt quickly spent the money.

Here *aunt, spent* belong to $\overline{S_0^{sub}}$, *the money* belongs to *Common*, and *quickly* to *REST*. In the following sections, I consider ordering principles which are not strictly dependent on the accessibility history.

4.3.3.3 Deictic expressions and pronouns

Discourse entities may be a mutually accepted part of the current discourse situation and serve to frame it for both speakers, for example, those corresponding to *I, you, here, now.* They are, in Prince's terms, 'situationally evoked' (Prince 1980), and assumed by both speakers to be highly accessible. They are therefore inclined to be

less prominent than other elements of the utterance:

(4.60) $$\mathcal{S}_0^{sub} \prec_{ix} REST$$

for $\mathcal{S}_0^{sub} \in$ *Deictics*.

However, when there is a clash with the historical rules, the latter win, otherwise we would rule out:

(4.61) John asked for chocolate
but <you> asked for <chips>

In (4.61), assigning reduced prominence to *you* is blocked by Rule 4.55, since a structural comparison begins at the level of the semantic equivalent of *ask for*, and assigns increased prominence to those parts which do not correspond exactly.[4]

Algorithms to determine when a discourse entity may be appropriately referred to using a pronoun have been considered extensively in the past (eg. Dale 1988) and are not strictly within the scope of this work. As regards accent, pronouns are not only reduced semantically, but it would appear that lack of metrical prominence, combined often with phonological reduction, may positively aid the listener to identify the referents of these highly accessible discourse entities (eg. Fowler and Housum 1987).

The prominence rule for '$\prec_p$' may therefore be stated in a similar form as that for deictic expressions:

(4.62) $$\mathcal{S}_0^{sub} \prec_p REST$$

for $\mathcal{S}_0^{sub} \subset$ *Pronouns*. An illustration of this rule is the following:

[7] T1:SA:2013 (T)
(4.63) *A*: you say it's arriving on the [@:m]
C: it's arriving early morning

Both tokens of *it* must be deaccented. As argued in Section 4.3.3.2, to do otherwise would evoke unwanted presuppositions.

[4]The possibility of using local contour to mark the theme-rheme distinction in examples such as (4.61) is further considered in Section 4.3.4.

4.3.3.4 Inferrables

One criterion for discourse entities to belong to Prince's class of inferrables, is the possibility of referring to them using a definite description, notwithstanding the absence of any previous mention, for example:

(4.64) the bus was late
the driver blamed the fog

It has been noted (eg. Sanford and Garrod 1981) that inferrables are unlikely candidates for pronominalisation. I follow Sanford and Garrod in distinguishing a subset of inferrables, the *scenario-dependent entities*, which, once a scenario has become established, may be assumed to be particularly accessible. In Sanford and Garrod's account, these are represented as unfilled slots which come into being with the scenario.

The results reported in Brown (1983) would suggest that inferrables are less likely to be defocussed than recent explicitly mentioned discourse entities. However consider the following extract:

(4.65)
[4a] T1:SB
C: so if I (.) you know (.)the plane is full
when I get there
they're not going to put another one on

This is the first mention in the dialogue of the new scenario (arrival for checkin); nevertheless the location: *there*, and the airline personnel: *they*, are both defocussed. This would suggest that scenario-dependent entities may well be relatively less prominent. My main concern here is with inferrable discourse entities which are not pronouns. Try replacing the pronouns in (4.65):

(4.66)
C: so if I (.) you know (.)the plane is full
when I get to the airport
the airline's not going to put another one on

Now although possible, it is much less likely that the expressions corresponding to inferrable discourse entities: *the airport* and *the airline* will be deaccented.

The discourse model semantics presented in Section 4.3.1 lacks an explicit formulation of scenario. However, in the human-machine dialogue situation it is rea-

sonable to equate scenario with a database task. We might during a conversation, for example, have a succession of task-scenarios such as the following:

flight enquiry: journey A

fares enquiry

flight enquiry: journey B

connection enquiry

reservation

where journeys A and B are different. Different possibilities for a given journey are assumed to belong to the same scenario. It is then possible to define the inferrable entities for a given task, *TaskInferrables*, to be those discourse entities which are known to exist and which are unique for their types, given the task. For example, if the task specifies a journey with a particular departure place and destination, then the discourse entities corresponding to the flight, the arrival event, and the departure event are scenario dependent. Flight numbers, on the other hand, are not, since there may be many which correspond to the same journey definition.

The rule for (scenario-dependent) inferrables:

(4.67) $$IDS_{inf} \prec_{if} REST$$

for $IDS_{inf} \in$ *TaskInferrables*, affords inferrables low precedence. However to account for examples such as (4.66), this rule must be given relatively low priority, and may even be optional.

4.3.3.5 Informativeness

There remains the case of discourse entities which are inherently more or less interesting than others within a given structure, being more or less informative. To some extent, relative *informativeness* is an emergent property of the factors considered above, such as inferrability. Here I concentrate on the static, context-independent aspects of this relation, whereby semantic entities within an expression are considered relative to one another. As is the case with other prominence relations so far discussed, we are dealing with a partial ordering, which may or may not apply to

an arbitrary pair of discourse entities within an expression.

Rather than attempting to define a mechanism which may be supposed to apply in the case of both speaker and hearer, I offer three heuristics which appear appropriate to the data and domain covered in this work.

1. Semantic information which gets incorporated in the discourse model is relatively more informative than information that doesn't. In particular, those parts of a semantic expression which contribute towards the interpretation of its dynamic significance within the dialogue, such as propositional attitude operators, are relatively less informative than those which refer to discourse entities. Thus in the structure:

$$\left[\begin{array}{l} id : want1 \\ type : want \\ thetheme : speaker \\ thegoal : \left[\begin{array}{l} id : go1 \\ type : go \\ thedest : london \end{array}\right] \end{array}\right]$$

   —corresponding to the utterance: *I want to go to London*, WANT1 is relatively uninformative. Similar cases would be entities of type REPEAT in *Could you repeat the departure time*, and SAY in *Did you say after seven thirty.* However this heuristic must be refined to deal with cases where the description of the dialogue act includes the feature *repeated* : +, and where this repetition is possibly the effect of the interlocutor's failure to comprehend the dialogue significance of the act the first time round, for example:

   (4.68)
   A did you say at <five>
   C sorry
   A did you <say> at <five>

   The issue of repeated utterances has already been discussed in Section 3.4.

2. Structures within the discourse model that contribute to task-domain information are more informative than ones that don't. Thus ARRIVE $\prec_{iv}$ TIME:330. Discourse entities such as ARRIVE, which do not directly contribute task information, may be regarded in this respect as placeholders, with a similar effect to function words or case labels.

3. Given two structures A and B, if the information content of A is inferrable

from that of B, then B is more informative than A. Consider for example the phrase *London Heathrow*. Mention of the discourse entity HEATHROW is more informative than that of LONDON; it is common knowledge that Heathrow is an airport in London. Determining informativeness on such grounds relies on assumptions about mutual knowledge; an individual who had not visited London might not know this fact, for example; another might not know that the less well-known Stansted was an airport in London, in which case an ordering which refused to assign relative priority to LONDON or STANSTED would be appropriate. Nevertheless a speaker guilty of 'egocentrism' (Chafe 1974) because of an unwarranted assumption is unlikely to mislead the hearer on grounds of prosodic prominence alone. If the hearer perceives the relative prominence (which is not guaranteed when surface ordering follows focus ordering), he may simply draw the conclusion that this was a case of world knowledge that he should have been aware of, but wasn't.

The rules for relative informativeness are summarised in (4.69) and (4.70):

(4.69) $$IDS_{\overline{dm}} \prec_{iv} IDS_{dm} - IDS_{task} \prec_{iv} IDS_{task}$$

where IDS_{dm} and $IDS_{\overline{dm}}$ are those semantic ids explicitly represented in the discourse model, and their complement; and IDS_{task} are those discourse model entities which are explicitly tied to the task.

(4.70) $$\forall\, ID_i, ID_j : one\_to\_many(ID_i, ID_j) \qquad ID_i \prec_{iv} ID_j$$

Rule (4.69) can be illustrated by the sentence: *You want to fly to Manchester*, where WANT $\prec_{iv}$ FLY $\prec_{iv}$ MANCHESTER. Rule (4.70) is illustrated in the Cyprus dialogue:

(4.71) [38] T2:SA:1516 (T)
30 C y'know the trains at gatwick station

Here the reading $>$ *gatwick* $<$ *station* is preferred; in other contexts (eg, discussing London stations) the reverse might be the case.

4.3.3.6 Prioritising the prominence rules

Having partitioned issues of relative prominence according to the five relations $\prec_{ix}, \prec_h, \prec_{if}, \prec_p, \prec_{iv}$, it remains to determine how these principles combine. Since all the relations are partial orderings, and what is required is a partial ordering, the simplest approach would be to merge the relations into one, by renaming each as the simple prominence order $\prec$. However difficulties arise in cases where a given pair of discourse entities are assigned conflicting orderings according to different prominence principles. For example, if LONDON has been recently mentioned, but DEPART hasn't, then LONDON $\prec_h$ DEPART. However this conflicts with the informativeness principle, according to which DEPART $\prec_{iv}$ LONDON.

The non-historical ordering relations largely complement one another. Relative informativeness for example is related to assumptions about inferrability. By contrast, $\prec_h$ may overrule prominence ordering derived from these. Consider for example the sentence:

(4.72) ba843 leaves before ba125
and it leaves before ba220

If *it* corefers with *ba843*, then $\prec_h$ and $\prec_p$ make the same predictions. Assuming the simplified semantics:

$$\begin{bmatrix} type : leaves\_before \\ earlier : X \\ later : Y \end{bmatrix}$$

the parallel structures are identical except for the fillers of the *later* slot, which is therefore marked as more prominent by $\prec_h$. But this is compatible with $\prec_p$, which marks *it* as less prominent (by $\prec_p$). The result,

BA843 $\prec$ LEAVES_BEFORE $\prec$ BA220

is a simple combination of $\prec_h$ and $\prec_p$.

On the other hand, if *it* corefers with BA125, comparison of the two structures

results in disparity for the fillers of both *earlier* and *later* slots, giving:

$$REST \prec_h \{ \text{BA125}, \text{BA220} \}$$

Although this violates $\prec_p$, $\prec_h$ must be allowed to take priority, or there would be no way of distinguishing between coreference in parallel structures which observes the parallelism, and coreference which doesn't. A similar general argument may be made, that where there is conflict between $\prec_h$ and other ordering principles, $\prec_h$ wins. An independent reason for giving $\prec_h$ priority is that to a certain extent this relation indicates how the current semantic structure can be built, re-using where possible previous structures from within the accessibility history.

A prominence ordering incorporating $\prec_{ix}$, $\prec_p$, $\prec_{if}$, $\prec_{iv}$ and $\prec_h$ is therefore a partial ordering incorporating all their ordering relations.

I represent a partial ordering in *normal form* as a set of ordered pairs, with those removed that can be derived by transitivity. The left-associative operator '$\oplus$' is defined to be set union over such partial ordering, with precedence to the left: $\prec_1 \oplus \prec_2$ consists of those pairs in $\prec_1$, with the addition of any pairs in $\prec_2$ which do not contradict relations in $\prec_1$ or its transitive closure. The overall prominence ordering, $\prec$ is therefore:

(4.73) $$\prec_h \oplus \prec_{ix} \oplus \prec_p \oplus \prec_{if} \oplus \prec_{iv}$$

The implementation of prominence ordering is discussed in Section 5.5.2.

The combination rules can be applied to the Agent's utterances, in the dialogue of Figure 4.47.

Firstly, for A1:*what time do you want to arrive in Paris*:

$$\frac{\{\text{PARIS}, \text{CALLER}\} \prec_h REST \quad \text{WANT} \prec_{iv} REST}{\{\text{CALLER}, \text{PARIS}\} \prec \text{WANT} \prec \{\text{TIME}, \text{ARRIVE}\}}$$

In the case of A3:*what time do you want to leave London*, we have:

$$\frac{\{\text{TIME, CALLER}\} \prec_h \textit{REST} \prec_h \{\text{DEPART, LONDON}\} \quad \text{WANT} \prec_{iv} \text{DEPART} \prec_{iv} \text{LONDON}}{\{\text{CALLER, TIME}\} \prec \text{DEPART} \prec \text{LONDON}}$$

For A4:*sa123 leaves London at twelve thirty*:

$$\frac{\{\text{DEPART, LONDON}\} \prec_h \{\text{SA123, 1230}\} \quad \text{DEPART} \prec_{iv} \{\text{LONDON, SA123, 1230}\}}{\{\text{DEPART}\} \prec \{\text{LONDON}\} \prec \{\text{SA123, 1230}\}}$$

Finally, for A4':*and arrives in Paris at two thirty*:

$$\frac{\{\text{SA123}\} \prec_h \textit{REST} \prec_h \{\text{ARRIVE, PARIS, 230}\} \quad \text{ARRIVE} \prec_{iv} \{\text{LONDON, SA123, 1230}\}}{\{\text{SA123}\} \prec_h \{\text{ARRIVE, PARIS, 230}\}}$$

4.3.4 Contrastive focus

In the analysis so far, prominence patterns indicating accessibility have been restricted to those where comparisons are between structures in a single, consistent information state. In this section I extend the model to cases of disparate information states, such as might arise in a repair situation or when alternative solutions are proposed. On the prosodic side, I assume that relative prominence may be supplemented by emphasis, and that additionally local contour accompanying a prominence marking give further indication about how the content relates to the discourse model.

4.3.4.1 Conflicting information states and repair

In this section I pursue the idea that contrastive focus comes about because of the need to highlight the difference between mutually incompatible information states. Consider the case where the Caller corrects an earlier statement:

(4.74)
[SA 3:C1] I was ringing to enquire about the flight SA 512 from paris. Is it on schedule
[SA 3:A2] 512 from paris
[SA 3:C2] sorry, 513
[SA 3:A3] 513 from paris

One reading of [SA 3:A3], in which the Agent takes note of the changed flight number, uses contrastive prominence to bring out the difference between the two

values:

(4.75) [SA 3:A3] 51≪3≫ from paris

This emphasis was indeed observed in all readings. The algorithm for deriving focus from the accessibility history would in fact give this pattern for [SA 3:C2], but not necessarily [SA 3:A3]. Moreover, being concerned with comparisons of a local, structural nature, it would fail to account for the observation that we are dealing with a repair, that conflicting values are at stake, and that the emphases, when made, are probably part of the speakers' intentions.

In [SA 3:C2] the speaker explicitly corrected himself—the discourse marker *sorry* explicitly drawing attention to the repair. Alternatively, the correction may be initiated by the other party:

	[SA 4:A8]	arrive rome eleven thirty
(4.76)	[SA 4:C8]	seven thirty
	[SA 4:A9]	no eleven thirty

The cases 4.74 and 4.76 correspond to Caller-initiated and Agent-initiated repair, respectively. In both cases, we may say that the speaker is *authorised* to make the repair, as the original *owner* of the information—ie, the first to introduce it into the discourse. On the other hand, [SA 4:C8] might be viewed as an unauthorised Caller-initiated repair; though it seems more likely that it is a case of simple mishearing. Consider also:

	[SA 8:C3]	is there a flight around mid morning
(4.77)		...
	[SA 8:A6]	well, there's a flight at eight fifteen

—where, it seems, A needs to stretch what counts as 'mid morning' to be able to retrieve a solution. This is a case of cooperative answering: where a solution that corresponds exactly to the Caller's requirements is lacking, a way out is to relax one or more of the initial constraints of the problem (eg. Kaplan 1983, Guyomard and Siroux 1989). Such a modification may be required by considerations of cooperativeness; it is nevertheless unauthorised, and its dispreferred nature may be expressed using a deferential tone, such as **HLH** , on *eight fifteen.* Further consideration of [SA 4:C8] yields another case of unauthorised repair, this time Caller-initiated. Suppose

that C did hear the time correctly, but the time *seven thirty* seemed unreasonable because it conflicted with his expectations about how long the journey should take. C's repair is not a correction, but calls a value into question, possibly with an emphatic contour conveying incredulity:

(4.78) ↑**HL$_\perp$H** seven thirty

There seems to be a difference between this fall-rise and the deferential contour that accompanies a cooperative response: here the overall range is wider, and the low point is more likely to reach the bottom of the speaker's range.

Another case of modification, which hardly counts as repair, is denial of a default value:

(4.79)
[SA 4:A2] are you travelling from Heathrow
[SA 4:C2] no Stansted
[SA 4:A3] travelling from Stansted

[SA 4:C2] is authorised, the Caller being the 'expert' about where he wants to travel from. But it would be unusual for *Stansted* in [SA 4:A3] to receive any contrastive prominence, because default values are readily overturned. On the other hand, an agent may resist the overturning of strong default assumptions, as in Example 4.78, or in the following:

(4.80)
A there's a flight from Clapham
C from ↑**HL$_\perp$H** Clapham

the incredulity stemming from the strong default assumption that Clapham—a relatively built-up area of south London—has no airport. A similar case is Example 4.27, reprinted here

(4.27)
[2a] T1:SA: (T)
C: well he's got four forty five leaving malaga
A: (well it's not our . . .)
H leaving ↑↑**HL** malaga

where the Agent is unable to find a flight corresponding to the Caller's specification, and so suggests the Caller may be mistaken.

To summarise, where there are inconsistencies between the information states of agents, utterances which initiate or confirm repairs serve to bring out the differences.

by highlighting them. This is a common factor behind the cases considered, whether or not the speaker has authority to make the repair; and whether a weak or a strong default has been overturned.

4.3.4.2 Representing conflicting information states

The cases of conflicting information states discussed above have been ones in which the conflict is local and temporary. Representing each information state by its own knowledge base of propositions is wasteful, and fails to bring out the locality of the discrepancy. Instead, the observation that speakers may introduce and refer to discrepancies via intonational emphasis suggests a representation based on differences. This is achieved using a layered representation of information states: each information state, or world, may inherit information from other worlds, so that only local additions and differences need to be represented. In particular, if information in a world W_i conflicts with that in a previous world W_{i-1}, the changed information is represented at W_i and nothing else.

Formally, let $W_0 \ldots W_n$ be a finite set of *world indices*: unique identifiers of information states. Define an inheritance relation *from_w* as a partial ordering over worlds, with W_0 distinguished as the initial world or root; every world index inherits at least from W_0. The local relation *parent_w* relating a world index and its parent is such that *from_w* can be derived as its transitive closure. I distinguish between local inheritance of two sorts: transparent and opaque. Transparent inheritance is monotonic—information can only be added at the daughter world, not taken away or modified. Opaque inheritance is non-monotonic. The relation *from_w* can thus be partitioned into *from_w*^{t} for transparent inheritance, and *from_w*^{o} for opaque inheritance. The knowledge local to a world index W_i is the set of propositions $inf\_local_{W_i}$ which W_i references. Then the total knowledge at a world index, inf_{W_i} can be defined recursively as follows:

(4.81) $$inf_{W_0} = inf\_local_{W_0}$$

(4.82) $$inf_{W_i} = inf\_local_{W_i} \uplus \bigcup_{W_j} inf_{W_j}$$

where *parent_w*(W_i, W_j). The definition states that the information at W_i consists of that defined locally, combined with that defined at ancestor worlds. The operator $\uplus$ is union, with overriding where the relation between the disparate worlds is an opaque one.

Semantic expressions which bring about a world with conflicting information, or refer to one, are conveniently represented as sets of substitutions, where each old element is replaced by some (different) new element. For utterances where the modification is fully explicit, both new and old parts are referred to, as in *not london, luton.* Alternatively it may be sufficient to refer just to the new part, relying on emphatic accent to identify it from within a larger matrix structure—[SA 3:C2. 3:A3].

I shall pursue the more general case of explicit modifications. These may be expressed semantically as instructions to replace the filler of a certain slot with a new value, for example, replace LONDON in < *departureplace* > with LUTON. A *modification* is defined to be a triple

$$\langle Ctx, Old, New \rangle$$

where *Old* is the value at *Ctx* in one world, and *New* is a different value at a different world. I define the *modifications* of a world W_i, *Mods*(W_i), as follows:

(4.83)
$$\begin{aligned} Mods(W_i) \triangleq \{ & \\ & \langle Ctx, Old, New \rangle \mid \\ & Old \tilde{\neq} New \wedge \\ & W_i \models value(Ctx, New) \wedge \\ & \exists\, W_j (from\_w^{o}(W_i, W_j) \wedge \\ & W_j \models value(Ctx, Old)) \\ & \} \end{aligned}$$

That is, *Mods*(W_i) enumerates the changes between W_i and some parent(s) W_j. The possibility that W_i may have more than one parent world is left open; however in most cases that I have examined, single inheritance of worlds seems adequate. In order to be efficient, a modification should be minimal. I shall take this to mean

that nowhere are any substructures of $Mods(W_i).old$ and $Mods(W_i).new$ identical, where

$$Mods(W_i).old \triangleq \{Old \mid \langle Ctx, Old, New\rangle \in Mods(W_i)\}$$

and $Mods(W_i).new$ is defined similarly. The requirement that two structures be recursively different is expressed by the relational operator '$\tilde{\neq}$'. This fits the cases described here, and is also in line with the computational treatment of modifications described in Section 5.2. It arguably lacks efficiency, since there will be cases where a number of modifications could be replaced by just one which breaks the recursively different constraint.

Having defined $Mods(W_i)$, the patterns of contrastive prominence for the cases:

- explicit modification: *not <u>five</u> thirty, <u>nine</u> thirty*;
- semi-explicit modification: *no, <u>nine</u> thirty*;
- implicit modification: *<u>nine</u> thirty*;
- denial: *not <u>five</u> thirty*

can be defined as follows:

(4.84) $EMod.1 : REST \prec\!\!\prec Mods(W_i).old$

(4.85) $EMod.2 : REST \prec\!\!\prec Mods(W_i).new$

(4.86) $IMod : REST \prec\!\!\prec Mods(W_i).new$

(4.87) $DEN : REST \prec\!\!\prec Mods(W_i).old$

where $EMod.1$, $EMod.2$ correspond to the first and second parts of an explicit modification expression, $IMod$ to an implicit or semi-explicit one, and DEN to a denial. Not surprisingly, $EMod.2$ and $IMod$ have the same prominence conditions, as do $EMod.1$ and DEN.

As an example, consider the case where the speaker wishes to modify the arrival time 530 to 930. Underlying this modification are two worlds with indices W_1, W_2:

$$W_1 \models value(<departuretime>, \begin{bmatrix} id : time\_point1 \\ type : time\_point \\ thehour : \begin{bmatrix} id : hour1 \\ type : hour \\ value : 5 \end{bmatrix} \\ theminutes : \begin{bmatrix} id : minutes1 \\ type : minutes \\ value : 30 \end{bmatrix} \end{bmatrix})$$

$$W_2 \models value(<departuretime>, \begin{bmatrix} id : time\_point1 \\ type : time\_point \\ thehour : \begin{bmatrix} id : hour2 \\ type : hour \\ value : 9 \end{bmatrix} \\ theminutes : \begin{bmatrix} id : minutes1 \\ type : minutes \\ value : 30 \end{bmatrix} \end{bmatrix})$$

The structures are shown in detail to demonstrate that they only disagree at the level of $<departuretime\ thehour>$; ie, the minimum modification is given by:

$$Mods(W_2) = \{(<departuretime\ thehour>, \begin{bmatrix} id : hour1 \\ type : hour \\ value : 5 \end{bmatrix}, \begin{bmatrix} id : hour2 \\ type : hour \\ value : 9 \end{bmatrix})\}$$

An explicit repair can be formed on the pattern: *not EMod*.1, *EMod*.2, where $W_1 \models EMod.1$, $W_2 \models EMod.2$. *EMod*.1 minimally will be $Mods(W_2).old$, and *EMod*.2 $Mods(W_2).new$, ie 5 and 9. They could however contain common material from W_1 or preceding worlds, as in *not leaving at 530, leaving at 930.* At least the minutes information is needed, to block the default 'oclock' interpretation. The pattern of prominences for *EMod*.1 is:

$$REST \prec\!\!\prec mods(W2).old$$

$$ie, \textsc{minutes1} \prec\!\!\prec \textsc{hour1}$$

and for *EMod*.2:

$$REST \prec\!\!\prec mods(W2).new$$

$$ie, \textsc{minutes2} \prec\!\!\prec \textsc{hour2}$$

Note that the identifier TIME_POINT1 may be dropped from the ordering, since it doesn't explicitly introduce a lexical entry.

A similar technique may be used to account for contrastive prominence where the roles differ. In the utterance:

(4.88) not <u>from</u> Heathrow, <u>to</u> Heathrow

This time the underlying worlds W_1 and W_2 give:

$$W_1 \models value(< departureplace >, \left[\begin{array}{l} \text{LOCATION1} \\ theairport : \text{HEATHROW} \end{array} \right])$$
$$W_2 \models value(< arrivalplace >, \left[\begin{array}{l} \text{LOCATION1} \\ theairport : \text{HEATHROW} \end{array} \right])$$

In the grammar used here, the words *from* and *to* are treated as semantically void; they serve as placeholders, indicating in the lexical entry for *travel*, for example, which subcategorial arguments fill which roles. The elliptical case (4.88) is difficult, because to associate the semantic roles with the prepositional placeholders, it is necessary to assume a matrix verb such as *travel*—corresponding to discourse entity GO1, then drop it. I assume that the word *travelling* is added, giving

(4.89) $$EMod.1 = \left[\begin{array}{l} id : go1 \\ type : go \\ departureplace : \left[\begin{array}{l} \text{LOCATION1} \\ theairport : \text{HEATHROW} \end{array} \right] \end{array} \right]$$

(4.90) $$EMod.2 = \left[\begin{array}{l} id : go1 \\ type : go \\ arrivalplace : \left[\begin{array}{l} \text{LOCATION1} \\ theairport : \text{HEATHROW} \end{array} \right] \end{array} \right]$$

The modification giving rise to the appropriate prominence pattern is:

$$Mods(W_2) = \{(< >, [\ departureplace : \text{HEATHROW}\ ], [\ arrivalplace : \text{HEATHROW}\ ]))\}$$

Then for the lexical entry *travelling*, the subcategorial argument corresponding to the *departureplace* role in (4.89)—ie, *from Heathrow*—is raised to contrastive prominence; inside this *Heathrow* can be demoted, giving:

(4.91) not travelling ≪from≫>Heathrow<

EMod.2 can be dealt with similarly.

In Section 5.2.1, I give details of how world indices are incorporated in the discourse model implementation. Section 5.3.3 explains how choices regarding contrastive prominence are made. Dialogues *nfenq*311, *nfenq*511, *nfenq*611 and *nfenq*811 in Appendix D illustrate the working of the Sundial message output sub-system, in various cases involving altered belief states.

4.3.4.3 Alternative solutions

The use of worlds to represent potentially disparate information states may be extended to representing multiple database solutions within the discourse model. A solution which satisfies a set of constraints supplied by the Caller may be represented within the discourse model as an instantiation of those constraints. For example, times which were loosely specified can be made precise with respect to a timetable. But simply adding the information of the solution has drawbacks:

1. The distinction between the original constraints and their satisfaction becomes blurred;

2. most important, there is no way of representing *multiple solutions*, each of which forms an alternative instantiation of the original constraints.

If the representation of the discourse model makes use of world-indices, a straightforward solution is possible. Given a set of constraints at world index W_c, together with n database solutions $Sol_1 \ldots Sol_n$, create n worlds $W_{c+1} \ldots W_{c+n}$, obeying:

$$(4.92) \qquad \forall i : 1 \leq i \leq n \qquad from\_w^t(W_{c+i}, W_c) \wedge inf\_local(W_{c+i}) = Sol_i$$

That is, a world is created for each distinct solution, inheriting transparently from the world where the original constraints were specified. As a result, the initial constraints are kept distinct from each solution, and any number of mutually incompatible solutions may be asserted, at different world indices.

For example, if the Caller gives the constraints

$$\left[\begin{array}{l} type : single\_journey \\ \\ thedeparture : \left[\begin{array}{l} \ldots \\ theplace : london \\ thetime : morning \\ thedate : monday \end{array}\right] \end{array}\right]$$

asserted at world index W_0, the alternative solutions:

$$\begin{aligned} <thedeparture\ thetime> &= 0840 \\ <thedeparture\ thetime> &= 0950 \\ <thedeparture\ thetime> &= 1020 \end{aligned}$$

may be respectively asserted at W_1, W_2, W_3. A sentence derivable from this information is given in (4.93):

(4.93) 6: A7 there are flights at 840, 950 and 1020

Consider now the more informative utterance in Example 4.94.

(4.94) **HLH** <BA123> leaves at <840>
HLH <BA226> leaves at <950>
and **HLH** <BA358> leaves at <1020>

Note that the prominences are only grossly specified; it is likely that *bee ay* will be defocussed in the latter two phrases. Utterances like (4.94) are best analysed in terms of theme-rheme organisation (eg. Bolinger 1989: 389–391), where the theme, or sentential topic, is intonationally marked in each case with a fall-rise.

Theme-rheme organisation as a unifying device has been explored by de Fries (1983), where it is applied to the *method of development* of a span of text, reflecting some principle of organisation such as spatial, temporal, or list organisation. In this study I concentrate on theme-rheme organisation as applied to parallel utterances, deriving from a set of database solutions. Here each solution is presented from the point of view of a local topic, or theme; in (4.94) this is the flight number. Theme-rheme organisation in an utterance may be based on the Caller's requirements, for example, when asked:

(4.95) when are there flights from london airports leaving in the morning

an appropriate reply might be

(4.96) from **HLH** Stansted at 840
from **HLH** Gatwick at 950
and from **HLH** Heathrow at 1020

Or the organisation may not explicitly reflect anything in the preceding discourse, but be based on a default model representing how information can be most conveniently packaged. This might be the case in (4.94); theme-rheme ordering might even reflect known characteristics of a database model, such as the physical layout of columns in a printed timetable.

Theme-rheme organisation may be represented by assuming that the hearer is being presented with information in the form of a functional relation over every

solution instance, with the themes corresponding to the domain of the function. Thus for (4.94), the function (4.97) can be defined, with an instance given by (4.98). This replaces the (less directed) relational pairs (4.99).

(4.97) $times\_for\_fids : FID \rightarrow TIME$

(4.98) $\{stansted \mapsto 840, gatwick \mapsto 950, heathrow \mapsto 1020\}$

(4.99) $\{\langle stansted, 840\rangle, \langle gatwick, 950\rangle, \langle heathrow, 1020\rangle\}$

I do not pursue how such a function is derived, but assume it can be dynamically defined for a set of solutions, either by virtue of the Caller's orientation, as in (4.95), or because of a default pattern of organisation, as in (4.94).

To display such organisation, an utterance must have a number of properties:

1. its components need to be structurally parallel;
2. theme-rheme organisation must be present across the components, so that in every component the theme and rheme are identifiable, and structurally parallel to corresponding themes and rhemes in other components;
3. if possible, theme and rheme are focally prominent, and identified apart intonationally.

The requirement that theme and rheme be focally prominent might be satisfied by founding this prominence on the difference between solution world indices and their parent, as was done for contrastive repair utterances in Section 4.3.4.1. However consideration of the following example suggests that prominences for theme and rheme may alternatively be derived via the accessibility history:

(4.100) from **HLH** Stansted at 840
from **HLH** Gatwick at 950
and from Gatwick at 1120

If the theme-rheme organisation is most important, then the second mention of *Gatwick* will have a **HLH** accent. But a better reading would probably make use of the accessibility history and defocus the second *Gatwick*.

Put formally, the function $\mathcal{TR}$ is defined over *Sol_Ids*, where

$$Sol\_Ids = \bigcup_{W_i} inf\_local(W_i).ids$$
$$(c+1 \leq i \leq c+n)$$

is the set of all solution entities; it partitions *Sol_Ids* exhaustively. Every solution is a relational instance of $\mathcal{TR}$:

$$\forall\, W_i (inf\_local(W_i) \subset \mathcal{TR})$$

I shall refer to the instance of $\mathcal{TR}$ particular to W_i as $\mathcal{TR}_{W_i}$. A further possible constraint is that all the themes belong to the same context, $theme(\mathcal{TR})$, and similarly, all rhemes belong to $rheme(\mathcal{TR})$:

$$\forall\, W_i \quad ($$
$$W_i \models value(theme(\mathcal{TR}), \mathrm{dom}\, \mathcal{TR}_{W_i}) \wedge$$
$$W_i \models value(rheme(\mathcal{TR}), \mathrm{ran}\, \mathcal{TR}_{W_i})$$
$$)$$

where dom and ran are operators on a function which return its domain and range respectively.

This function does not directly affect prominence ordering for each component of the utterance. Under normal conditions, where solutions are all distinct, the historical accessibility prominence ordering $\prec_h$ (cf. Section 4.3.3.2) will give prominence to all themes and rhemes. Thus for Example 4.100, $\mathcal{TR}$ consists of the relation instances:

$\langle$STANSTED, 840$\rangle$
$\langle$GATWICK, 950$\rangle$
$\langle$HEATHROW, 1120$\rangle$

If two successive rhemes are the same, the second will be given reduced prominence by $\prec_h$:

from **HLH** Stansted at 840
(4.101) from **HLH** Gatwick at >840<
...

Two themes may also be the same, violating the constraint that $\mathcal{TR}$ be a function:

from **HLH** Stansted at 840
(4.102) from **HLH** Stansted at 920
...

Example 4.102 suggests that we relax the functional constraint, which isn't really necessary, and require only that $\mathcal{TR}$ be a many-to-many mapping, for which the themes and the rhemes are contextually aligned. But the equally plausible reading:

from **HLH** Stansted at 840
(4.103) from >Stansted< at 920
...

in which *Stansted* is demoted because of $\prec_h$, could be allowed by permitting the pair $\langle$STANSTED, 920$\rangle$ to be excluded from $\mathcal{TR}$, or even allowing all the entries corresponding to STANSTED to be grouped together, so that the relational instance becomes:

$$\langle \textit{stansted}, \langle 840, 920, \ldots \rangle \rangle$$

I take the second solution, of dropping from $\mathcal{TR}$ relation instances that would violate functionality.

The function does however impose a linearisation constraint whereby if possible the parts of a phrase corresponding to the theme precede those corresponding to the rheme. Surface order constraints in the lexicon may or may not allow theme-rheme ordering to go ahead; if necessary, where lexical choice exists, entries which allow this ordering should be preferred. If the ordering is lexically blocked, it may still be imposed via some stylistic ordering device such as topicalisation, or passivisation. On the other hand, intonational marking of $\mathcal{TR}$ instances is sufficiently strong for the linearisation constraint to be overruled. Consider Example 2.43 (reprinted here):

(2.43) a) London's the capital of Scotland isn't it?
b) No **HL** Edinburgh's the capital of **HLH** Scotland]
HLH London's the capital of **HL** England.

The functional mapping which would naturally emerge from a knowledge representation (say a geographical database) would be that of (4.104), with the instance (4.105):

(4.104) $country\_capital : COUNTRY \rightarrow CITY$
(4.105) $\{england \mapsto london, scotland \mapsto edinburgh\}$

In place of this straightforward mapping, (2.43b) presents a 'skewed' one

(4.106) $\{scotland \mapsto edinburgh, london \mapsto england\}$.

Here the domain and range are of heterogeneous types; this arises out of the need to repair a previous misconception.

In Section 5.3.3 I give details of the implementation of the theme-rheme relation.

4.3.5 Discourse-model-related focus: a summary

If we assume that the relative prominence in surface trees, as manifest for example in prosodic accent, derives in part from relative prominence at a conceptual level, this semantic prominence may come about in a number of ways. Prominence is inversely related to accessibility. Semantic structures may be accessible because they are the same, or modified versions, of recent structures. Deictic expressions such as *I, you, here, now* are also taken to be particularly accessible. Discourse entities which for particular reasons not considered in this thesis are represented as pronouns, are treated as highly accessible and hence of low prominence. Discourse entities not explicitly introduced, but inferrable, are potentially higher in the prominence hierarchy than highly accessible entities. Within a semantic structure, components which are more interesting on context-independent grounds (such as those which reference task information) will by default be more prominent. All of these orderings may be overruled by prominence derived from modification with respect to the accessibility

history. Even more prominent (generally raised to emphatic prominence) are discourse entities corresponding to modifications with respect to a former information state.

Discourse entities in parallel structures resulting from describing alternative solutions to a query are ordered for prominence according to the historical/modification principles. However the information may be further organised by giving it the direction of functionality, manifest both textually and prosodically in the theme-rheme structure of the utterance.

4.4 Focus assignment and speaker's intention

I have shown (Section 4.3.4) that when a repair utterance needs to address the fact that an information state has become modified, the focus properties of the message that convey the modification are integral to the speaker's intention. This may be contrasted with the situation where focal properties merely serve to aid the listener by conveying accessibility information (Section 4.3.3.2). In the latter case, utterances with inappropriate focus patterns will, *ceteribus paribus*, still succeed in their intentions. This is not the case for repair utterances. Consider the utterance [SA 4:A9] uttered in the following manner

(4.107) <eleven><thirty>

If this is preceded with *no*, the hearer may infer (incorrectly) that minutes, or indeed the whole time expression, are in need of modification; without any such discourse marker, there is no indication that this is a repair utterance.

The observation that repair utterances are necessarily specified for certain patterns of focus, may be generalised to dialogue acts in general:

(4.108) A dialogue act which is referential (ie, not a phatic utterance) may require for its successful performance that certain focal elements be made prominent, the nature of the prominence depending on the dialogue act.

I shall refer to this as the *principle of intended focus.* Thus confirmation utterances require that the material needing confirmation be prominent; similarly utterances

conveying new information require it to be prominent. The intented focus principle is important for a number of reasons. It represents another ordering principle, which generally takes priority over those based on accessibility. Often the two mechanisms will make similar predictions. Where the intention is to present new information, for example, the accessibility mechanisms will result in that information being raised in prominence, because accessible information is generally lowered. However the accessibility mechanism will not always deliver appropriate results: take the case of a confirmation request, consisting of a repetition. The accessibility mechanisms are incapable of singling out the element(s) requiring confirmation. The intentional focus of the utterance must include discourse entities about which doubt exists. If it is *Stansted* that requires confirmation in *from Stansted on monday*, for example, raising the discourse entity STANSTED in prominence will result in an appropriate ranking:

(4.109) from ≪Stansted≫on monday

—where *monday* is assumed not to require explicit confirmation.[5]

As in Section 4.3.4, I shall assume that local contours may be exploited to more precisely indicate the nature of the focus.

4.4.1 Intended focus in the case of repair utterances

In order to characterise intended focus more exactly, I return to the case of repair utterances which make use of contrastive focus. A repair act incorporating a modification requires that attention be drawn to the modification. This means that in the case of an implicit repair with contrastive focus on $Mods(W_i).new$ (cf. Section 4.3.4.2), making these items prominent is part of the intention of the utterance.

Consider now the various situations in which modification utterances can be made:

Confirming an authorised modification: Consider again Example 4.74:

[5] By virtue of its nuclear position, with other informational elements in the tail, it is not necessary in this case for the accent on *Stansted* to be emphatic.

(4.74)
[SA 3:C1] I was ringing to enquire about the flight SA 512 from Paris. Is it on schedule
[SA 3:A2] 512 from Paris
[SA 3:C2] sorry, 513
[SA 3:A3] 513 from Paris

In [SA 3:A3], *A* is confirming an authorised repair. Two contours seem possible:

(4.110) five one$^{\uparrow}$**HLH** three from Paris
(4.111) five one$^{\uparrow}$**HL** three from Paris $^{\downarrow}$]

The emphatic prominence on 'three' is shown as a boosted pitch accent; however emphasis could equally well be conveyed by amplitude or phrasing. Apart from the emphasis, the major difference is the nuclear tone. In (4.110) there is a clearer indication that the speaker expects the confirmation utterance to be taken as strong. Example 4.111 may still elicit a response; however unlike (4.110) a *yes* answer is probably expected. This makes it, in the terminology of Brown *et al.* (1980), a *conducive* question. In the Swedish Corpus, all speakers in fact used a falling nucleus.

Confirming doubtful information: Take again (4.110), uttered this time in a context not where a modification is being made, but where *A* has failed to hear clearly one of the digits of *C*'s preceding utterance. In this case it is possible to use a discourse representation based on the disparity between two worlds, W_1 and W_2, where

$$\begin{aligned} W_1 &\models value(<flightnumber>, \text{BA}\,51?) \\ W_2 &\models value(<flightnumber>, \text{BA}\,513) \\ & \quad parent\_w(W_2, W_1) \end{aligned}$$

and W_2 is marked to indicate that information local to it is of doubtful validity. Once again, the modified element is emphasised; however on this occasion, the **HLH** contour is more appropriate: a **HL** contour would misleadingly suggest confidence in the value.

Authorised correction: The Caller has uttered a value incompatible with that previously provided by the Agent, and which the Agent is expert about, as in (4.76)—reproduced here:

(4.76)
[SA 4:A8] arrive Rome eleven thirty
[SA 4:C8] seven thirty
[SA 4:A9] no eleven thirty

In 4:A9, *A* is authorised to make the change, since *C*'s supposed confirmation initiative involves an unauthorised modification. The most likely reading (observed in the recordings) is

(4.112) no e↑**HL** leven thirty

An **HLH** contour would be possible here, but tends to sound deferential or patronising (compare *unauthorised modification* below). In the Swedish Airlines corpus, speakers used the **HL** contour.

Unauthorised modification: The Agent may wish to modify the constraints set by the Caller, in order to avoid the even more dispreferred situation of not responding at all (eg. Kaplan 1983, Guyomard and Siroux 1989). Thus in Example 4.77 (reproduced here)

(4.77)
[SA 8:C3] is there a flight around mid morning
...
[SA 8:A6] well, there's a flight at eight fifteen

eight fifteen is probably a weakening of the constraint *mid morning.*

There seems to be a strong association in English between such constraint-relaxation utterances, and the use of the fall-rise. For example, [SA 8:A6] may be uttered

there's a flight at **H** eight fif**HLH** teen

Ladd (1980) and Ward and Hirschberg (1985) attempt to give truth-conditional accounts, in their treatment of similar utterances (cf. Section 2.5.4). Instead

I treat this as a distinct component of a suggestive act intended to convey a placatory attitude on the part of the speaker. Evidence for this is the use of this contour when an agent is pleading, cajoling, or in some way attempting to persuade. It then seems more reasonable to invoke pragmatic principles, to the effect that when an agent is unable to fulfil some assumed requirement of the interlocutor, (s)he is socially required to be deferential. That set-theoretic accounts of the use of the fall-rise accent are misleading may be demonstrated by the following example:

(4.113)

10 am:	*Nick*:	There are ten sweets in this bag Don't eat all of them
10.30 am:	*Nick*: *Ben*:	Well, I bet you did eat all the sweets I ate one

In (4.113), Ben would be unlikely to feel the need to be deferential, since he was not violating any of Nick's expectations. However, an emphatic fall on *one*, conveying indignation, would be appropriate.

Note furthermore, that (4.77) still works when the fall-rise on *eight fifteen* is replaced with a fall; however, the dispreferred nature of the act is no longer softened.

Confirmation of a modification overturning defaults: We have seen (Section 4.3.4.1) that the nature of prominence on the modified item depends on the strength of the default that was overturned:

(4.79)

[SA 4:A2]	are you travelling from heathrow
[SA 4:C2]	no stansted
[SA 4:A3]	travelling from stansted

(4.80)

A	there's a flight from Clapham
C	from ↑$\mathbf{HL_{\perp}H}$ Clapham

whereas in (4.79)—reproduced here—contrastive prominence on *stansted* is not appropriate, in (4.80) contrastive prominence on *Clapham*, together with a tone conveying incredulity, is acceptable.

4.4.2 Intended focus in non-repair utterances

A repair utterance which refers to a modification of information requires that some form of prominence be given to items which refer to that modification. In the case of non-repair utterances with informative content, it may still be necessary to give this content prominence. I consider the two cases of (non-repair) confirmations, and informative utterances.

Confirmations: I suggest above that confirmation requests where some modification has taken place, or where there is serious doubt about some value, be treated together with repair utterances, as cases requiring contrastive prominence. However, it may be argued that we have a continuum, whereby prominence is given to elements to the extent that they are in need of confirmation. If for example e_1, e_2, e_3 are the elements requiring confirmation (in increasing order of need), the prominence requirements may be stated:

$$REST \prec e_1 \prec e_2 \prec e_3$$

If there is confidence about a value, it may be relatively or absolutely defocussed. This makes it considerably more difficult for the interlocutor to effect a repair if the value is wrong:

(4.114)
A1: travelling to **HL** Paris from >London<
C1: from ↑**HL**$_{\perp}$**H** Luton

Informative utterances: The rules for prioritising accessibility-related prominence (cf. Section 4.3.3.6) already rank informative entities concerning details such as times and flight numbers over mere placeholders such as TRAVEL, ARRIVE. The relation '$\prec_{iv}$' carries with it no guarantee of special prominence; however an utterance intended to impart information must do so by ensuring that the information is maximally intelligible. Giving prosodic prominence to the new information achieves this. Thus in

(4.115)
A5: there are flights at **HLH** seven, eight fif**HLH** teen,
and **H** ten ↓**HL** thirty

from Dialogue 1 of the Swedish Corpus, prominence is required on *seven, eight fifteen* and *ten thirty.* A distinction can be made between the prominence requirements here, and those for repair utterances, which I have shown to mostly require emphatic prominence. The use of emphasis in (4.115) is pragmatically ruled out, since it would presuppose a modification—unless, that is, the new information does violate some existing constraint. As far as contours are concerned, Example 4.115 clearly demonstrates that these are derived at the phrase level, not at the level of sentence accents.

In both cases of non-repair utterances discussed, the intentional component of focus is relatively weak; emphasis and contours with local scope are not required.

4.4.3 Representing intended focus

Intended focus means that certain discourse entities referred to in an utterance be given special prominence, as part of the speaker's intention. I shall call the set of such entities the *information profile* of the message underlying the utterance. Under normal conditions, if an utterance has a non-empty information profile, this acts as a lower limit on possible ellipsis; ie, elliptical utterances must at least reference all elements of the information profile. I have shown that elements of the information profile are typically more prominent than others in the utterance, the nature of the prominence depending on the intention of the utterance. In addition, in certain cases a local contour (or tonal accent) which tends to accompany utterances of that intention will be associated with accents corresponding to information profile elements. In Table 4.4, I summarise the kinds of prominence. Typical tonal accents are given, where these are specific to a particular intention, together with references to the SA corpus. The degenerate case where no special prominence is associated with an element requiring confirmation is also dealt with, by associating no prominence type.

Association of intended focus with a message $\mathcal{M}$ can then be described as follows: If $\mathcal{M}$ is to be described using an ordered set of discourse entities $\langle ids\_\mathcal{M}, \prec_{\mathcal{M}} \rangle$, where $ids_{\mathcal{M}}$ are the discourse entities and $\prec_{\mathcal{M}}$ the instance of the prominence relation that

orders them, then the required focus specification $focus(\mathcal{M})$ is the structure

$$\langle ids\_\mathcal{M}, ip_{\mathcal{M}}, \prec'_{\mathcal{M}}, tone_{\mathcal{M}} \rangle$$

where $ip_{\mathcal{M}} (\subseteq ids\_\mathcal{M})$ is the information profile corresponding to $\mathcal{M}$; $\prec'_{\mathcal{M}}$ is derived from $\prec_{\mathcal{M}}$ by applying the associated prominence type to members of $ip_{\mathcal{M}}$, overriding where necessary; and $tone_{\mathcal{M}}$ is the tonal accent associated with $\mathcal{M}$, where specified. In the intonational contour that results, constituents corresponding to members of $ip_{\mathcal{M}}$ will receive the appropriate degree of prominence, as well as being associated with $tone_{\mathcal{M}}$. Details of contour derivation in the implemented system are presented in Chapter 5.

4.5 Towards a unified account of utterance production

4.5.1 Focus assignment as a part of utterance production

In the analysis of focus-assignment mechanisms up till now, computational principles underlying the production of utterances have been largely assumed. However an account of focus assignment which shares mechanisms with those required independently for the production of utterances, if it can be achieved, is to be preferred. In this section I give an overview of the computational mechanisms that are assumed to be required for language production, indicating how the focus assignment principles proposed fit in with these mechanisms.

intention	prominence type	tonal accent	corpus
confirm authorised modification	$\prec\prec$	$^{\uparrow}$**HL**	[SA 3:A3]
confirm doubtful information	$\prec\prec$	**HLH**	
authorised modification	$\prec\prec$	($^{\uparrow}$)**HL**	[SA 4:A9]
unauthorised modification	$\prec\prec$	**HLH**	[SA 9:A4]
weak default overturned	$\prec$	**HL**	[SA 4:C2]
strong default overturned	$\prec\prec$	$\uparrow$**HL**$_{\perp}$**H**	Ex. 4.80
confirmation	$\prec$ or none		
informative	$\prec$		

Table 4.4: Association between intention and effects on prominence

Production starts with an intention to speak. In the model of dialogue proposed in Section 3.3, this intention takes the form of a dialogue act label combined with a reference to the informational elements that must be included in the message. The latter are equivalent to the information profile of the message, already discussed in Section 4.4.3. Starting with this input, production down to the level of a prosodically structured text string, takes place in three major phases. Firstly, a directed acyclic graph (DAG) which spans and may extend the information profile, is derived as a subgraph of the discourse model. Secondly, this graph may be extended by semantic entities which lie outside the domain of the discourse model, as defined here: namely, those indicating propositional attitude components or serving as discourse markers and connectives. This is done on the basis of the dialogue act label, which may contain information about the role of the act in the greater conversational context. Additionally syntactic features indicating properties of the utterance such as mood, and other stylistic preferences, are added. Finally, an analysis tree is built, using knowledge from the lexicon to derive grammatical and lexical constraints which are added to the existing semantic ones. Interleaved with this last process, or as a post process, assignment of prosodic prominence and contours takes place. In Chapter 5 I discuss in more detail the design and implementation of the first stage (description building: Section 5.3) and the third stage (linguistic generation: Section 5.4).

4.5.2 A unified account of focus assignment

There appears to be a certain amount of redundancy between the mechanisms for assignment of focal prominence operating at the linguistic, attentional and intentional levels. Take for example, the case of contrastive focus in a repair utterance. The fact that the utterance will encode a modification is already present at the intentional level, where it may even be indicated what the modification is to be. At the attentional level, the modification is apparent from the discrepancy between information states; again at the linguistic level, providing the replaced structure has appeared in a recent utterance, the modification can be constructed using the linguistic history. Consider again the case of defocussing relatively accessible material.

At the attentional level, the recent occurrence of the relevant discourse entities will be apparent from the accessibility history; but this information is also present, in the form of reusable linguistic structures, in the linguistic history. This suggests that, while the analysis at the individual levels has appeared reasonable, something must be done at a global architectural level both to eliminate this redundancy from the account of processing, and to ensure that the different levels do not throw up contradictory results. As I argue below, there are a number of reasons why this apparent superfluity should be tolerated:

1. Some phenomena can only be explained at a certain level. If we reconsider the example discussed on Page 135, an attempt to draw structural parallels between the conceptual representations NINE SEVEN and ONE SEVEN will be ignored at the lexical level, since *seventeen* doesn't decompose, and will therefore fail to be capable of being marked as parallel to the surface structure *ninety seven*. On the other hand, it might seem reasonable to argue that numbers are not decomposed into their digits at a conceptual level, this happening only in the lexicon. In this case, the account of focus with respect to the linguistic history is required, to make sense of parallel structures which can be prosodically marked, such as *thirty seven* and *forty seven*. Similar issues are raised by considerations of phrases incorporating the word *double*, such as occur in some utterances of telephone numbers. Apart from this, prominence of affixes, as in *not de*ˈ*duce, re*ˈ*duce* can only be accounted for at a coarse level by the conceptual accessibility mechanisms, and needs in addition comparison of syllables at a surface level.

   It must be also admitted that a mechanisms proposed at some level may be more ungainly than one achieving the same thing at a different level. Take contrastively stressed prepositions, as in *not from luton, to luton* Because of cases where *from luton* was never explicitly said, as in the following:

(4.116)
C: I want to travel to luton. I need to leave at five
A: leaving luton at five
C: not from luton, to luton

I proposed in Section 4.3.4 an approach at a conceptual level to modifications which involve changed roles rather than changed values. But a surface structural approach is much more straightforward.

2. The different levels may complement each other, each contributing to the result. Compare the cases of responses to violation of weak and strong defaults. Both involve at the attentional level discrepancies between the information states. But the prosodic marking of the two cases tends to be quite different, the changed items being played down in the weak case, and exaggerated, maybe with a tone conveying incredulity, in the strong case. The incredulous response is clearly inviting a reaction from the interlocutor; the toned-down response can be relatively indifferent. Such a decision is best taken at the intentional level.

3. a 'levels' approach is still valuable even if all the decisions on the speakers part are taken at a single level, because in using the prosodic signals, the hearer may be in an asymmetrical position *vis-à-vis* the speaker. The decision taken by the speaker at the conceptual level to indicate reuse with respect to the accessibility history may be used most efficiently during parsing by the listener.

4. independently of considerations of focus assignment, the temporary configurations at one level of production may help at a later level, if they can be fed forward. For example, even admitting that a linguistic history mechanism is useful for dealing with structural phenomena not accountable for at a conceptual level, most instances of reuse which are detected in the accessibility history could go forward to the linguistic history and eliminate the search that would otherwise be necessary to retrieve the relevant surface forms.

The implementation described in Chapter 5 takes this last point into account, by allowing accessibility signals produced at the description stage to feed forward to the linguistic generation stage, where they guide the surface reuse algorithm.

4.6 Summary

In this chapter I have considered mechanisms at different levels which may be said to govern the assignment of prominence during production. The levels may be termed *linguistic, attentional* and *intentional.* At the linguistic level the possibility of accessing previous surface structures, as stored in the linguistic history, was discussed. At the same time, a representation was introduced capable of describing both lexical entries, and, by extension, analysis trees. At the attentional level, I introduced the concept of the discourse model, and defined a representation for it. Various criteria, according to which a discourse model may be considered to be more or less accessible, were presented. In particular, an accessibility history which records previous conceptual structures forming part of the discourse was proposed. I introduced the notion of information states indexed by worlds in order to handle cases of discrepancies, and alternative solutions. At the same time, it was possible to relate these to the repair utterances which often come with such discrepancies. At the intentional level, I introduced the information profile, which the message can be thought to be about. This necessitates special prominence for discourse entities which are part of it.

Finally, I indicated how the mechanisms for prominence assignment proposed in this chapter may be combined within a full account of utterance production by the speaker. The mechanisms may be mutually redundant, or they may conflict. I therefore presented a framework within which the mechanisms can be seen to support one another, and make the correct decisions at the appropriate levels.